12-2019

Model Selection for Longitudinal Data With Time-Dependent Covariates Using Generalized Method of Moments

Maryann Nishimura Shane

UNIVERSITY OF NORTHERN COLORADO

Greeley, Colorado

The Graduate School

MODEL SELECTION FOR LONGITUDINAL DATA WITH TIME-DEPENDENT COVARIATES USING GENERALIZED METHOD OF MOMENTS

A Dissertation Submitted in Partial Fulfillment
of the Requirements of the Degree of
Doctor of Philosophy

Maryann Nishimura Shane

College of Education and Behavioral Sciences
School of Applied Statistics
and Research Methods

December 2019

This Dissertation by: Maryann Nishimura Shane

Entitled: *Model Selection for Longitudinal Data with Time-Dependent Covariates Using Generalized Method of Moments*

has been approved as meeting the requirement for the Degree of Doctor of Philosophy in College of Education and Behavioral Sciences in School of Applied Statistics and Research Methods, Program of Applied Statistics

Accepted by the Doctoral Committee

_____
Trent Lalonde, Ph.D., Research Advisor

_____
Han Yu, Ph.D., Committee Member

_____
Khalil Shafie, Ph.D., Committee Member

_____
James Kole, Ph.D., Faculty Representative

Date of Dissertation Defense _____

Accepted by the Graduate School

_____
Cindy Wesley, Ph.D.
Interim Associate Provost and Dean
The Graduate School and International Admissions

ABSTRACT

Shane, Maryann Nishimura *Model Selection for Longitudinal Data with Time-Dependent Covariates Using Generalized Method of Moments.* Published Doctor of Philosophy dissertation, University of Northern Colorado, 2019.

The purpose of this dissertation was to establish measures that could be used to assess the relative fit of nested models with parameters estimated using the Generalized Method of Moments for longitudinal data with time-dependent covariates. A secondary data set collected from Filipino children was used as an example of model fitting to evaluate the quality of the assessment of fit of the Kullback-Leibler Information Criterion (KLIC) and a chi-squared statistic derived from the difference in the minimums of the quadratic forms of two candidate nested models. A simulation involving randomly-generated data sets was also used to evaluate the performance of the proposed statistics. Several variations of nested models were considered in the simulation, and the KLIC was used to compare the relative fit of these models.

Overall, the performance of the KLIC as a model selection criterion showed that it achieved good detection proportion in identifying the correct model when it was compared to underfit models. On the contrary, it tended to favor overfit models over the correct model, and non-detection proportions were high when extraneous predictors were introduced to candidate models. Ignoring the feedback loop introduced by time-varying covariates and relying on the regular use of the Generalized Estimating Equations (GEE) for the analysis of longitudinal data could compromise model parameter consistency,

efficiency, and bias resulting in misleading inferences. Replacing the former practice with the routine use of GMM to properly account for feedback in the data is highly encouraged. The KLIC would be a helpful tool to select an appropriate model among a collection of candidate GMM models, especially when there are time-varying predictors in the data.

Keywords: model selection, fit statistic, information criterion, Generalized Method of Moments, time-dependent covariates, longitudinal data

ACKNOWLEDGEMENTS

This research would not have been possible without all the support I received. I dedicate this dissertation to my parents: to my mother, the most patient and caring person whose love always gets me through the hardest times; to my father, who always believed I would make something of my life and pushed me beyond my abilities. I want to thank my sister, Tina Shane, for encouraging me and making me believe in myself, and my partner in crime, Garrett Banks, for teaching me to see things through to the end.

To my friends Karen Traxler and Ben Overholt, who constantly reached out to me throughout the dissertation process, reminding me that I had someone to lean on if I ever needed help with anything: thank you both for your friendship and support. To my mentors Ron Filadelfo and Jim Jondrow: thank you for your endless support with my personal and professional development--you believed in me from Day 1, took me under your wings, and helped me become a better person through leading by example.

To my adviser, Trent Lalonde, for guiding me through a long and windy road; this research could not have happened without your support. I would also like to thank my committee members, other friends from the Department of Applied Statistics and Research Methods, my editor Judieth Hillman, and anyone else not included in this acknowledgement who have encouraged me throughout this dissertation. An astounding number of people in my life helped me in different ways and cheered for my success throughout this research. This dissertation would not have been possible without each and every one of you. Thank you all for your endless encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Change is an inevitable phenomenon in the natural and physical world. The concept of change has motivated human beings to study science to better understand the underlying elements that drive it. The Baltimore Longitudinal Study of Aging, conducted by the National Institute of Health, has been one of the lengthiest studies on human aging. Its objective was to understand the process of aging and the biological changes associated with aging, as well as behavioral, genetic, and external factors that impact these changes.

In the analysis of independently observed data, classical Maximum Likelihood estimation has been the most common approach. Maximum Likelihood requires knowledge of the response distribution and employs the use of the likelihood function. Maximum Likelihood estimation has oftentimes been used to estimate canonical parameters due to its property to always yield consistent estimators; moreover, if unbiased, they often have minimum variance of all unbiased estimators (McCullagh & Nelder, 1989; Mendenhall, Scheaffer, & Wackerly, 1981; Wackerly, Mendenhall, & Scheaffer, 2008).

In longitudinal studies, data are repeatedly collected from the same subjects over time. These types of studies have been frequently used in medical, educational, and environmental practices to assess the impact of a treatment or intervention over time. Compared to using the more traditional cross-sectional studies, in which observations

are made on subjects at a single point in time, longitudinal designs have the ability to detect change over time using repeated measurements of the same variables on the same subjects. Zeger and Liang (1992) support the use of longitudinal designs to analyze and better understand change over time, emphasizing their two major advantages: robustness to model selection and increased power, due to increased sample size and subjects serving as their own baseline (Zeger & Liang, 1992).

However, with more complicated designs come more complex issues involving the analysis of such data. Unlike the methods available for handling independently observed data, such as maximum likelihood, the correlation inherent in longitudinal data must be properly accounted for during analysis to prevent consequences involving modeling issues, including bias, inconsistency, and inefficiency of parameter estimates (Diggle, Heagerty, Liang, & Zeger, 2002; Fitzmaurice, 1995; Fitzmaurice, Laird, & Ware, 2011; Hedeker & Gibbons, 2006; Pepe & Anderson, 1994).

In the analysis of longitudinal data, marginal and conditional models are available. In the marginal model, the modeling of the mean and the within-subject correlation are conducted separately (Zeger & Liang, 1992), whereas the conditional model--also commonly known as the random-effects model--conditions the average response on the covariates and additional variables (Diggle et al., 2002). Marginal models focus on modeling the population average of the response, the marginal mean, whereas conditional models focus on the assumption of a level of homogeneity of repeated observations on the same subject and heterogeneity across different subjects (Diggle et al., 2002; Fitzmaurice et al., 2011). As a result, marginal models involve population-averaged conclusions, while conditional models yield subject-specific conclusions.

One of the most popular marginal methods of obtaining model parameter estimates for correlated data has been the Generalized Estimating Equations (GEE) approach (Liang & Zeger, 1986; Zeger & Liang, 1986), which employs the use of the quasi-likelihood of the response, rather than the full likelihood, as assumed in classical maximum likelihood estimation. The quasi-likelihood requires specification of only the mean and mean-variance relationship rather than the full specification of the response distribution. Similar to the popular Akaike Information Criterion, or AIC (Akaike, 1973, 1974), an information-based measure used to assess the global fit of models constructed using a likelihood-based approach, the quasi-likelihood information criterion, QIC, and its adjustment, QICu, have been used to assess the overall fit of models obtained using a quasi-likelihood approach (Pan, 2001a).

Generalized Estimating Equations (GEE) has been widely used in the handling of longitudinal data in many disciplines; however, it has encountered challenges in the presence of covariates that introduce a time-dependency to the data structure. These variables, known as time-dependent covariates (TDCs), could compromise model parameter consistency and efficiency, sometimes introduce bias when longitudinal data with TDCs are modeled using GEE, and may result in misleading inferences about the parameter estimates (Diggle et al., 2002; Fitzmaurice, 1995; Fitzmaurice et al., 2011; Pepe & Anderson, 1994). To evade some of these issues posed by GEE, the Generalized Method of Moments (GMM) has been proposed for statistical model building of correlated data (Hansen, 1982, 2007; Lai & Small, 2007). Rather than employing the use of the quasi-likelihood of the model parameters as with GEE, the algorithm underlying

GMM uses moment conditions with zero expectation to build a quadratic form that is minimized over all parameters for model construction (Hansen, 1982).

Although GMM may improve the efficiency of model parameter estimates in the presence of TDCs (Lai & Small, 2007), one of its major disadvantages has been the lack of a universally accepted fit statistic for model selection. This research study presents two measures that could be used to compare the fit of nested models using a moment-based estimation procedure in the presence of TDCs. The focus of the first fit statistic was on an information criterion, similar in application to the popular Akaike Information Criterion (AIC; Akaike, 1973), and could be used to compare multiple nested models for the purpose of variable selection in the presence of time-dependent covariates. The second fit statistic was based on the existing quadratic form from GMM: the difference in the quadratic forms of two candidate models fit using GMM could be used to compare them inferentially using hypothesis tests.

The Purpose and Focus of the Study

The purpose of this study was to establish measures that could be used to assess the relative fit of nested models with parameters estimated using the Generalized Method of Moments for longitudinal data with time-dependent covariates. The current literature was sparse in its discussion of variable selection for correlated models involving time-dependent covariates. Model selection is important from a general modeling standpoint, and it is especially critical when different nested models require the use of different amounts and levels of resources, including time, cost, and manpower.

The focus of this study was on establishing measures to assess model fit using a moment-based method. For independently observed data, classical estimation procedures,

such as maximum likelihood, are used to construct models in which the overall goodness-of-fit or relative fit of nested models could be evaluated using statistics like the Akaike Information Criterion (AIC), model deviance, and the Bayesian Information Criterion (BIC). In using the Generalized Estimating Equations (GEE) to build models for correlated response data, the Quasi-likelihood Information Criterion (QIC) and its adjustment, QICu, could be used to assess model fit. However, when model parameters are estimated using an approach that does not involve the likelihood or quasi-likelihood functions, there was no universally accepted statistic that was used to assess model fit and compare nested models. The goal of this study was to establish such a statistic; moreover, focus was placed on an information-based measure that is analogous to the more common AIC and QIC information-based statistics. Further, a second statistic that follows a known distribution was presented, allowing researchers to compare the fit of two specific models. The process is similar to the comparison of two candidate models using the model deviance, in which the difference of the log-likelihoods of the two models is distributed as a chi-squared statistic with degrees of freedom equal to the difference in the number of parameters of the two models.

The Need for the Study

Measures to compare nested models fit using a moment-based estimation procedure with time-dependent covariates were investigated in this study. The current literature was sparse in its discussions of model fit involving the Generalized Method of Moments and other methods that do not involve the likelihood or quasi-likelihood functions; moreover, the literature lacked discussions about the selection of time-dependent covariates. After its establishment, the Akaike Information Criterion (Akaike,

1973, 1974) has been used popularly by researchers when applying likelihood-based estimation posed with the issue of model selection. New statistics to assess GMM model fit should benefit researchers interested in moment-based estimation techniques in obtaining model parameter estimates for correlated data with time-dependent covariates.

With the era of information technology expansion and availability of bigger and more extensive data, more research has been conducted using longitudinal designs to understand change over time. However, there was very little discussion of model selection for longitudinal data involving time-dependent covariates in the current literature, which placed a strain on the credibility of inference made by analysts who rely on the use of moment-based estimation for building correlated response models. This research was necessary in order to address the issue of model selection involving candidate nested models that are constructed using a moment-based procedure, such as the Generalized Method of Moments, when time-dependent covariates are present.

The Rationale for the Study

The statistics proposed in this research filled a gap in the current literature pertaining to the practice of model selection for longitudinal designs with time-dependent covariates. Model selection has been important to statistical model building, and debate has continued over the traditional methods and appropriate measures for model selection. When model-based inference is at stake, the dispute over which statistic to use--even with sophisticated measures, such as the Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC)--appeared ceaseless (Buckland, Burnham, & Augustin, 1997; Burnham & Anderson, 2004; Chaurasia & Harel, 2013). Inferential results are

valid only if the selected model meets all the necessary assumptions underlying appropriate statistical modeling practices.

When multiple predictor variables are available to model the outcome of interest, and hence, several nested models are under consideration, one of the key factors that influences the decision to select a given model is the parsimony of the model relative to the amount of information lost from the data in modeling the response outcome (Burnham & Anderson, 2004). This concern has historically been addressed with techniques involving the use of information-based criteria for model selection (e.g., AIC, BIC, QIC, Kullback-Leibler divergence principle, etc.).

Until this research, there has been no universal statistic used to assess the fit of nested models constructed using the Generalized Method of Moments in the presence of time-dependent covariates. Measures were presented in this research study addressing the issue involving moment-based model selection.

Research Questions

The research questions answered in this study were:

Q1 How can information associated with the fit of model parameters estimated using the Generalized Method of Moments be expressed or measured?

Q2 What is the detection proportion of the model selection process of such measures in their ability to detect poor fit of underfit models?

Q3 What are the non-detection proportions of such measures in suggesting poor fit for appropriate models?

Overview of the Methodology

In order to evaluate the quality of the assessment of fit of the proposed statistics, a secondary data set collected from Filipino children to understand the relationship

between nutrition and diarrheal diseases (Bhargava, 1994; Bouis & Haddad, 1990) was used as an example of model fitting. A simulation involving randomly-generated data sets was also used to evaluate the performance of the proposed statistics. Several variations of nested models were considered, and the proposed statistics were constructed to compare the relative fit of these models.

The simulation was not the main focus of the methods presented in this dissertation research. Its purpose was to allow the simulation of additional data sets for the sake of evaluating the performance of the proposed statistics in assessing model fit.

Example Data Analysis

Researchers of the International Food Policy Research Institute (IFPRI) collected data from Filipino children, aged 1-10 years, from the island of Mindanao between 1984 and 1985 to study the relationship between nutrition and health (Bhargava, 1994; Bouis & Haddad, 1990). Age, gender, height, weight, food consumed, illnesses suffered, the duration of illnesses, as well as descriptive information about the parents, were collected from 448 households using 4 surveys each 4 months apart. These primary data were used as secondary data in this research study. As correlation in nutritional quality was expected from siblings or from children within the same household, data from only the youngest child were kept. Additionally, observations with missing values were omitted, resulting in balanced longitudinal data containing information from 370 children at 3 time points (Bhargava, 1994; Bouis & Haddad, 1990).

From these longitudinal data, five variables were selected as potential predictors, as well as a binary response, to estimate a logistic regression model. In addition, a variation of the response defined by a transformation of the illness-related variables was

used as a continuous response variable for a multiple linear regression model. The information-based measure to assess fit was obtained and compared for all possible models that could be constructed from different combinations of five predictors. This measure was used to determine the most “ideal” candidate model to predict the probability of illness for the logistic regression model and to predict morbidity for the multiple linear regression models. Then, these models were compared to those obtained in the study by Lai and Small (2007).

Using the models that yielded the five smallest values of the information criterion, the statistic based on the difference in the minimums of two GMM quadratic forms from pairs of candidate models was obtained to assess significant departure of the candidate model from a model with adequate fit. The model selected as most ideal using the information-based measure of fit was compared to the remaining four models, for a total of four pairwise comparisons.

The Proposed Simulation

As only one set of models could be constructed from the real data, only one set of statistics could be obtained to assess the quality of fit of the candidate models for each of the two proposed statistics. Therefore, there was a need for a simulation to generate additional data sets for the sake of evaluating the performance of the proposed information criterion. A simulation was used to obtain more estimates of the two proposed statistics--the information-based statistic and the measure based on the minimums of two GMM quadratic forms--to assess their performance in analyzing fit. The software environment R version 3.1.0 was used to produce these data and perform all necessary analysis.

For the data simulation, both binary and continuous correlated responses were randomly generated to estimate logistic regression and multiple linear regression models. Continuous predictors, including time-dependent covariates (TDCs) of Types II and III, were also randomly generated, to keep the data structure consistent with that of the Filipino child study. A true model was defined as one that included three predictors, two of which were TDCs of Types II and III. The third predictor in the true model was a Type I TDC, to simplify the data analysis procedure. Two unnecessary predictors that were TDCs of Types II and III were also randomly generated. With attention paid to the performance of the proposed fit statistics when TDCs of Types II and III were improperly included or omitted, two overfit and two underfit models were examined in comparison to the true model. One of the overfit models included an additional unnecessary Type II TDC, and the other overfit model included an unnecessary Type III TDC. These models were used to assess the non-detection proportions of the proposed fit statistics, or the proportion of times in the simulation that an incorrect predictor was not detected by the KLIC. Similarly, one of the underfit models omitted a necessary Type II TDC, and the second underfit model omitted a necessary Type III TDC, and these models were used to evaluate the detection proportion of the proposed statistics in assessing adequate model fit, or the proportion of times that an incorrect predictor was correctly detected by the KLIC.

In the simulation, 2000 data sets (Lai & Small, 2007) were randomly generated for both small-sample and large-sample situations, as well as for the binary and continuous response cases. Following the simulation study by Lai and Small (2007), the small sample included a total of 500 observations, and the large sample included a total

of 2,500 observations. For these data to be balanced in longitudinal structure with T = 5 repeated observations per subject, the small sample included I = 100 subjects, and the large sample included I = 500 subjects. A discussion of the issues presented by the use of unbalanced data was included in the Limitations section in Chapter V. For each of the sample size conditions, the proposed fit statistics were obtained for each replicate.

The organization of this research follows: Chapter II provides an extensive review of the current literature involving the subject matter, and Chapter III delineates the methods used in evaluating the performance of the proposed statistics in assessing model fit. The analysis of the real data and the details of the simulation are described in Chapter IV, and a discussion of the results and conclusions that could be drawn from this research study are included in Chapter V.

CHAPTER II

REVIEW OF THE LITERATURE

In the literature review, various fit statistics were discussed for different estimation methods of parameters for longitudinal data models, with a focus on logistic regression models. For classical maximum likelihood estimation of independent data, the most common statistics used in assessing model fit was the deviance, the Akaike Information Criterion (AIC), and the Bayes Information Criterion (BIC), also known as the Schwarz Information Criterion (Cavanaugh & Neath, 1999). When using Generalized Estimating Equations (GEE) to estimate model parameters for correlated data, modifications of the AIC, known as the quasi-likelihood information criterion (QIC) and its adjustment, QICu, were used. There has been no popular method yet in assessing goodness-of-fit (GOF) for models derived from the Generalized Method of Moments (GMM) process. A review of the literature highlighted the need for such a GOF statistic, and details are given in the subsequent chapter.

Linear Regression

A linear regression model assumes the equation:

$$Y_{it} = \beta_0 + \beta_1 x_{it_1} + \beta_2 x_{it_2} + \ldots + \beta_k x_{it_k} + \varepsilon_{it},$$

where i = 1,..., N denotes subjects; t = 1,..., T denotes observation time; the mean response $Y_{it} = g^{-1}(\beta_0 + \beta_1 x_{it_1} + \beta_2 x_{it_2} + \ldots + \beta_k x_{it_k})$ for subject i is a function of the parameters ($\beta_0, \beta_1, \ldots, \beta_k$); g is the identity-link function; x_{it_j} is the j[th] covariate value at

time *t* for subject *i* for j = (*1,…, k*), and ε_{it} is random error. For independent data, the *i* index is maintained to identify subjects, but there is no time index (*t*), denoting different observation times.

Logistic Regression

The primary focus was on a logistic regression model:

$$\log it(p_{it}) = \ln\left(p_{it} \big/ 1 - p_{it}\right) = \beta_0 + \beta_1 x_{it_1} + \beta_2 x_{it_2} + \ldots + \beta_k x_{it_k} \ , \quad (1)$$

where *i* = 1,..., N denotes subjects; *t* = 1,..., T denotes observation time; the mean response $p_{it} = g^{-1}(\beta_0 + \beta_1 x_{it_1} + \beta_2 x_{it_2} + \ldots + \beta_k x_{it_k})$ for subject *i* is a function of the parameters ($\beta_0, \beta_1, \ldots, \beta_k$); *g* is the logit-link function; and x_{it_j} is the j[th] covariate value at time *t* for subject *i* for j = (*1,…, k*).

For independent data, the *t* index in Equation (1), denoting different observation times, was omitted; the *i* index was maintained to identify subjects. In matrix notation, the logistic regression model in Equation (1) could be written as:

$$g(\boldsymbol{p_{it}}) = X_i^T \boldsymbol{\beta} \ , \quad (2)$$

where *g* is the logit-link function; $\boldsymbol{p_i}$ is the vector of mean responses for subject *i*, which is equivalent to the probability of success for binary responses; $\boldsymbol{X_i} = \left[1, x_{it_1}, x_{it_2}, \ldots, x_{it_k}\right]^T$ is the vector of covariates for subject *i*; and $\boldsymbol{\beta} = \left[\beta_0, \beta_1, \ldots, \beta_k\right]^T$ is the vector of parameters. Additionally, a binomial response distribution was assumed:

$$Y_{it} = Bin(n, p_{it})$$

As with ordinary regression models, logistic regression was used to model the mean response--in this case, the probability of success. The conditional mean, $E\left[Y_{it} \mid \boldsymbol{X_i}\right]$,

was modeled, rather than modeling the expected response, $E[Y_{it}]$ (Hosmer & Lemeshow, 2000):

$$E[Y_{it} \mid \boldsymbol{X}_i] = \beta_0 + \beta_1 x_{it_1} + \beta_2 x_{it_2} + \ldots + \beta_k x_{it_k} . \quad (3)$$

A link function was used to link p_{it}, the probability of success, with the parameters of the model. In the case of a logistic regression model, the logit-link function, $\text{logit}(p_{it}) = \ln\left(p_{it} / 1 - p_{it}\right)$, is commonly used to link p_{it} with the model parameters (Hosmer & Lemeshow, 2000; Hosmer, Lemeshow, & Sturdivant, 2013).

Longitudinal Data Models

Longitudinal data, by definition, are data that are collected from the same individuals, or objects, multiple times using the same measure(s). Examples of longitudinal studies included the National Health and Nutrition Examination Survey (NHANES) and the National Education Longitudinal Study (NELS). The NHANES is an annual survey research study conducted by the National Center for Health Statistics (NCHS) to assess the health and nutritional status of Americans between the ages of 1 and 74. Fifteen participants have been followed annually since 1999 to maintain repeated observations of the same subjects, and the longitudinal database focuses on repeated measurements from these 15 individuals. National Education Longitudinal Study (NELS) is an education survey conducted by the National Center for Education Statistics (NCES) to assess the achievement and learning of eighth graders. A sample of the original group of eighth graders was followed four times at irregular intervals since 1988, and the main focus of these data was on the longitudinal information obtained from this subgroup.

In the handling of longitudinal data, the assumption of independently and identically distributed (i.i.d.) observations is not maintained due to the correlation inherent in repeated observations of subjects. As a result, an estimation method that accounts for correlated data, a process that could detect changes in the mean response over time, must be employed. To do this, two types of models were commonly used in the existing literature: the marginal model and the conditional model. For the sake of this research, a balanced design was assumed. A balanced design is one in which the responses of every subject are observed an equal number of times at equal intervals apart. A logistic regression model was still considered, as described at the beginning of this chapter.

The Marginal Model

In the marginal model, the mean and the within-subject correlation are modeled separately (Zeger & Liang, 1992). The marginal mean is the average response, conditioned on the covariates of subject *i* (Diggle et al., 2002), as represented by Equation (3). The marginal model links the marginal mean to covariates via the relation in Equation (2) using a known link function, *g* (Diggle et al., 2002; Fitzmaurice, 1995; Fitzmaurice et al., 2011; Hedeker & Gibbons, 2006; Zeger & Liang, 1992). As the focus of this research was in dealing with binary responses, *g* was the logit-link function.

In the marginal model, the focus is on modeling the within-subject correlation separately from the marginal regression of the response on the predictor variables (Diggle et al., 2002). In other words, the primary focus of marginal models is to model the systematic variation associated with the mean separately from the random variation that arises as a result of the repeated observations (Fitzmaurice et al., 2011). Assumptions are

made about the mean response, $E[\boldsymbol{Y}_i]$, as well as the covariance of the responses, $Var(\boldsymbol{Y}_i)$, and estimates are obtained for both the vector of parameters, $\boldsymbol{\beta}$, and the vector of within-subject correlations, $\boldsymbol{\alpha}$ (Diggle et al., 2002). Additionally, the marginal variance is a function of the marginal mean such that:

$$Var(Y_{it}) = v(p_{it})\varphi,$$

where v is a known function and φ is the over-dispersion parameter which accounts for the variation in Y_{it} not explained by $v(p_{it})$. The covariance between Y_{it} and Y_{is}, for a time point s ≠ t, is a function of the marginal means and possibly additional parameters $\boldsymbol{\alpha}$,

$$Cov(Y_{is}, Y_{it}) = c(p_{is}, p_{it}; \boldsymbol{\alpha}),$$

where c is a known function and the values of $\boldsymbol{\alpha}$ are the parameters in the correlation or covariance matrix. In a marginal model, the correlation between two repeated observations from the same subject is assumed to depend only on the time between the two measurements, represented by $\boldsymbol{\alpha}$ (Zeger & Liang, 1986).

Marginal models yield population-averaged conclusions (Zeger, Liang, & Albert, 1988). For logistic regression scenarios, the results of population-averaged models involve comparisons between two populations, or a reference group and a compared group, using odds ratios. Comparisons between the populations are expressed as the average change in the expected transformed response for a unit change in the value of a predictor of interest, holding all other predictors constant (Hosmer & Lemeshow, 2000; Hosmer et al., 2013; Menard, 2002).

The Conditional Model

In contrast to the marginal mean, the conditional mean is the average response conditioned on the covariates $\boldsymbol{X_i}$ and additional variables $\boldsymbol{B_i}$:

$$E\left[Y_{it} \mid \boldsymbol{X_i}, \boldsymbol{B_i}\right] = \beta_0 + \beta_1 x_{it_1} + \beta_2 x_{it_2} + \ldots + \beta_k x_{it_k} + \boldsymbol{B_i}.$$

The conditional model links the conditional mean to both the covariates $\boldsymbol{X_i}$ and the additional variables $\boldsymbol{B_i}$ using the conditional regression parameters $\boldsymbol{\beta}$ and additional parameters $\boldsymbol{\gamma}$ (Diggle et al., 2002; Fitzmaurice et al., 2011; Hedeker & Gibbons, 2006):

$$\boldsymbol{g}\left(\boldsymbol{p_i}\right) = \boldsymbol{X_i}^T \boldsymbol{\beta} + \gamma^T \boldsymbol{B_i} \ .$$

This is akin to a random-effects model, in which the correlation among responses for a given subject is assumed to arise from natural heterogeneity in regression coefficients across different individuals (Zeger & Liang, 1986, 1992). The additional variables, $\boldsymbol{B_i}$, are assumed to contain information about this homogeneity within and heterogeneity across subjects: the conditional model assumes that there are unobserved factors underlying the homogeneity of responses within a subject, thus inducing correlation within repeated observations on the same subjects, but that those factors vary across different individuals (Diggle et al., 2002).

Conclusions drawn from conditional models differ from the population-averaged conclusions made from marginal models. Conditional models are associated with subject-specific conclusions (Zeger et al., 1988). Parameter coefficients of conditional models are interpreted as the average change in the expected transformed response associated with a change in the predictor variable for a specific subject, holding all other predictors constant. However, conditional models were not the primary focus of this research; the marginal model was used in this study.

Time-Dependent Covariates

Time-dependent covariates (TDCs) are predictors whose values change over time, within a group, subject, or cluster. Weight, height, age, and study cohort are all variables

that could be TDCs. Any covariate that changes in value over time or changes over the course of repeated observations could be a time-dependent covariate (Diggle et al., 2002; Lai & Small, 2007; Neuhaus & Kalbfleisch, 1998). The age of a participant in a longitudinal study to assess hypertension is an example of a time-dependent covariate because the participant's age would increase over time. Annual glacial coverage in an ongoing study to assess causation and impact of global climate change, as well as the amount of chemical substance present in the half-life of a radioactive material, are also examples of time-dependent covariates. These special types of covariates introduce correlation among variables over time, and this correlation must be accounted for when constructing models for longitudinal data. Failure to account for this correlation when constructing models may result in loss of efficiency and increase the possibility of biased parameter estimates (Fitzmaurice, 1995; Lai & Small, 2007; Pepe & Anderson, 1994).

The importance of accounting for the temporal nature of time-dependent covariates and the subsequent impact on the analysis of longitudinal data is highlighted by the difference between endogenous and exogenous covariate processes. The following section differentiates these processes and provides an explanation in the context of a data feedback loop.

Exogenous and Endogenous Covariate Processes

Diggle et al. (2002) defined an exogenous process as one in which the covariate at a given time, *t*, was conditionally independent of response measurements prior to that time. In other words, a covariate process is exogenous if there is no response feedback to the covariate. Exogeneity implies that the mean response for subject *i* at time *t*,

conditioned on all covariate values (at other times) x_{i1}, x_{i2},..., x_{iT}, only depends on the covariates prior to time *t*:

$$E\left[Y_{it} \mid x_{i1}, x_{i2}, ..., x_{it}\right] = E\left[Y_{it} \mid x_{i1}, x_{i2}, ..., x_{i,t-1}\right].$$

Exogeneity also implies that the response for subject *i* at time *t* is conditionally independent of all future covariates, given both the past response and covariate values (Diggle et al., 2002). Any variable external to a study is an exogenous covariate. In the previously mentioned example of a study in which age was examined to assess a participant's risk of hypertension, weather was a possible exogenous variable.

On the other hand, an endogenous process is one in which feedback may be present: the response for subject *i* at time *t* could be associated with the covariate value at future time points. Diggle et al. (2002) described an endogenous covariate as both a predictor of the outcome of interest, as well as a measure that was predicted by the outcome at an earlier time. Hence, an endogenous process could involve a complex feedback loop in which the covariate influences the response, and the response influences the covariate (Diggle et al., 2002; Zeger & Liang, 1991).

Types of Time-Dependent Covariates

Recent literature has defined four types of time-dependent covariates (TDCs), and distinctions were made based on the nature of their feedback (Lai & Small, 2007; Lalonde, Wilson, & Yin, 2014). Covariate types were defined as follows. Consider a model defined as:

$$\boldsymbol{p}_i = g\left(\boldsymbol{X}_i^{T}\,\boldsymbol{\beta}\right),$$

where $\boldsymbol{p_i}$ is the mean response for subject *i*, *g* is a known link function, $\boldsymbol{X_i}$ is the vector of covariates for subject *i*, and $\boldsymbol{\beta}$ is the vector of model parameters. The four types of TDCs

were defined by the combinations of *s* and *t* that maintain the equality in the expression (Lai & Small, 2007; Lalonde et al., 2014):

$$E\left[\frac{\partial p_{is}(\beta_0)}{\partial \beta_j}\{y_{it} - p_{it}(\beta_0)\}\right] = 0, \quad (4)$$

where p_{is} is the mean response for subject *i* at time *s*, β_j is the j^{th} covariate, y_{it} is the response for subject *i* at time *t*, p_{it} is the mean response for subject *i* at time *t*, β_0 is the vector of true parameters, and *s* and *t* are different observation times, where $s \in (1,\ldots,T)$ and $t \in (1,\ldots,T)$.

A Type I time-dependent covariate satisfies the zero expectation of Equation (4) for all values of *s* and *t*, $s \in (1, \ldots, T)$ and $t \in (1, \ldots, T)$. Type II TDCs satisfy the zero expectation of Equation (4) for all combinations of *s* and *t* such that $s \geq t$. Type III TDCs satisfy Equation (4) for $s = t$. Lastly, a Type IV TDC satisfies the zero expectation of Equation (4) for all combinations of *s* and *t* such that $s \leq t$. Association between the derivative term at time *s* and the residual term at time *t* violates the equality in Equation (4), resulting in a non-zero expectation. In order for the zero expectation to be held, the assumption:

$$E\left[y_{it} \mid \boldsymbol{X}_{it}\right] = E\left[y_{it} \mid \boldsymbol{X}_{i1},\ldots,\boldsymbol{X}_{iT}\right],$$

must be maintained for all *s* and *t*, for $s \in (1, \ldots, T)$ and $t \in (1, \ldots, T)$. When time-varying covariates are present in the data, this assumption is oftentimes violated. Details are discussed in the Generalized Estimating Equations and Time-Dependent Covariates section.

A Type I TDC involves no feedback; the current covariate affects only the current response. A covariate that affects both the current and future responses is a Type II TDC.

The nature of the feedback involved in the presence of Type III TDCs creates a complex feedback loop in which the current response affects the covariate at some future time, and the current covariate affects the response at some future time. Type IV TDCs are often thought of as the "opposite" of Type II TDCs; when Type IV TDCs are present, the current response is associated with the current covariate, and the current response could also be associated with the covariate at some future time (Lai & Small, 2007; Lalonde et al., 2014).

Due to the nature of the time-dependence imposed by TDCs on the data structure, the analysis of longitudinal data in the presence of TDCs could become challenging very quickly (Diggle et al., 2002; Neuhaus & Kalbfleisch, 1998). Moreover, depending on the method of analysis chosen, algorithm non-convergence of parameter estimation may be an additional hurdle in the process of obtaining model parameter estimates (Kleiber & Zeileis, 2008).

Generalized Estimating Equations

One of the most popular techniques for marginal model parameter estimation in the analysis of correlated data has been the Generalized Estimating Equations (GEE) approach (Liang & Zeger, 1986; Zeger & Liang, 1986). The GEE method uses an assumed working correlation structure for the data under investigation. A working correlation structure (e.g., compound symmetry, order-*m* auto-regressive, exponential, etc.) that likely characterizes the nature of the correlation prevalent among repeated response measurements for each subject is proposed. Generalized Estimating Equations (GEE) uses estimating equations of the form:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \boldsymbol{p_i}}{\partial \boldsymbol{\beta}} \text{cov}^{-1}(\boldsymbol{y_i})(\boldsymbol{y_i} - \boldsymbol{p_i}) = 0,$$

where $\boldsymbol{p_i}$ is the vector of mean responses for subject *i*, $\boldsymbol{\beta}$ is the vector of parameters, and $\boldsymbol{y_i}$ is the vector of observed responses for subject *i*, for a total of *N* subjects (Liang & Zeger, 1986; Zeger & Liang, 1986). The covariance term, $cov(\boldsymbol{y_i})$, is calculated using the working correlation structure.

An advantage of using Generalized Estimating Equations in the analysis of longitudinal data could be that, regardless of whether or not the "correct" working correlation structure is selected, GEE parameter estimates are consistent (Diggle et al., 2002; Fitzmaurice, 1995; Liang & Zeger, 1986; Neuhaus & Kalbfleisch, 1998). Additionally, the GEE approach does not require full specification of the response distribution. It only requires information involving the mean and mean-variance relationship of the response, hence only assuming a quasi-likelihood instead of the full likelihood (Liang & Zeger, 1986; Zeger & Liang, 1986).

The covariance matrix used in the GEE process depends on the selection of the working correlation structure, $\boldsymbol{R_i}(\boldsymbol{\alpha})$ (Fitzmaurice, 1995; Hin & Wang, 2009; Liang, Zeger, & Qaqish, 1992; Zeger & Liang, 1986), where the parameters $\boldsymbol{\alpha}$ are part of the structure of $\boldsymbol{R_i}$. If the compound symmetry structure is assumed, then the correlation $\alpha_{s,t}$ among pairs of time points *s* and *t* is equivalent regardless of the combination of times; hence, only one value (i.e., α) need be estimated. The estimating equations for the vector $\boldsymbol{\alpha}$ have the form:

$$S(\alpha)=\sum_{i=1}^{N}\frac{\partial \eta_i}{\partial \alpha}\mathrm{cov}^{-1}(w_i)(w_i-\eta_i)=0\,,$$

where $\boldsymbol{\alpha}$ is the vector of parameters from the specified working correlation structure, $\boldsymbol{w_i}$ is the vector of covariances for pairs of responses at different combinations of time points *s*

and t (s, $t \in T$), and $\boldsymbol{\eta_i}$ is the vector of expected covariances, or $\boldsymbol{\eta_i} = \mathrm{E}[\boldsymbol{w_i}]$ (Liang et al., 1992).

In order to obtain the GEE estimators, the estimating equations would iteratively be solved for the regression coefficients, $\boldsymbol{\beta}$, and the correlation parameters, $\boldsymbol{\alpha}$. Given an estimate of the working correlation structure, $\boldsymbol{R}_i\left(\hat{\boldsymbol{\alpha}}\right)$, iteratively reweighted least squares is applied to obtain an updated $\hat{\beta}$ (McCullagh & Nelder, 1983). Once $\hat{\boldsymbol{\beta}}$ is obtained, a second set of estimating equations is used to obtain consistent estimates of $\boldsymbol{\alpha}$. This process is repeated until convergence is achieved (Liang et al., 1992).

Generalized Estimating Equations and Time-Dependent Covariates

Oftentimes in longitudinal designs, time-dependent covariates are present in the data structure. The Generalized Estimating Equations method was often applied to obtain parameter estimates in the presence of time-dependent covariates. Pepe and Anderson (1994) and Fitzmaurice (1995) advised that, when TDCs were present in the data, a critical assumption behind the Generalized Estimating Equations process should be confirmed when using it for parameter estimation. Specifically, GEE relies on the assumption of the marginal expectation,

$$E\left[Y_{it} \mid x_{it}\right] = E\left[Y_{it} \mid x_{is}, s = 1,...,T\right], \tag{5}$$

where Y_{it} is the response for subject i at time t ($t = 1, \ldots, T$), x_{it} is the covariate value for subject i at time t, and x_{is} is the covariate value for subject i at time s ($s = 1, \ldots, T$). This assumption states that for subject i, the expected response at time t, given the covariate value at that same time, should be equal to the expected response at time t, given the covariate value at *all* times (Fitzmaurice, 1995; Pepe & Anderson, 1994). In other words,

for a given subject, the covariate values at other times would not affect the conditional expected response at time *t*.

The marginal expectation represented by Equation (5) was an assumption that was necessary for the zero expectation of the Generalized Estimating Equations. As an alternative to checking this assumption, using a diagonal working correlation--such as the independent working correlation structure--ensures that the expectation of the generalized estimating equations is $\boldsymbol{0}$ (Diggle et al., 2002; Fitzmaurice, 1995; Pepe & Anderson, 1994):

$$E\left[S(\boldsymbol{\beta})\right]=E\left[\sum_{i=1}^{N}\frac{\partial \boldsymbol{p_i}}{\partial \boldsymbol{\beta}}\operatorname{cov}^{-1}\left(\boldsymbol{y_i}\right)\left(\boldsymbol{y_i}-\boldsymbol{p_i}\right)\right]=\mathbf{0}.$$

Evaluating the expectation is equivalent to integrating over all $\mathbb{R}^T$,

$$\Rightarrow\int_{\Re^T}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \boldsymbol{p_i}}{\partial \boldsymbol{\beta}}\operatorname{cov}^{-1}\left(\boldsymbol{y_i}\right)\left(\boldsymbol{y_i}-\boldsymbol{p_i}\right)\right)d\boldsymbol{y_i}=\mathbf{0}.$$

The only random components of $S(\boldsymbol{\beta})$ are the response values; the derivative and covariance matrices are composed of fixed values, so they could be brought outside of the integration process, provided that the derivative of the systematic component is independent of the raw residuals:

$$\Rightarrow\frac{1}{N}\sum_{i=1}^{N}\left(\int_{\Re^T}\frac{\partial p_i}{\partial \beta}\operatorname{cov}^{-1}\left(y_i\right)\left(y_i-p_i\right)\right)dy_i=0.$$

Using a diagonal covariance matrix such as the independent structure, $\boldsymbol{I^T}$, ensures that the derivative terms and residual terms at only the same times are matched, satisfying the assumption of independence for a correctly specified model (Diggle et al., 2002; Fitzmaurice, 1995; Pepe & Anderson, 1994). The use of a diagonal covariance matrix

maintains the zero expectation, and the assumption in Equation (5) is no longer relevant, and the integral simplifies to a scalar:

$$\Rightarrow \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial p_{it}}{\partial \beta_j}\int_{\Re^T}\left(y_{it}-p_{it}\right)dy_{it}=0\,.$$

However, if the covariance matrix is not diagonal, derivative and residual terms from various time points are matched, and the expectation may not equal $\boldsymbol{0}$ (Diggle et al., 2002; Pepe & Anderson, 1994).

Oftentimes when time-dependent covariates are present, the equality in Equation (5) does not hold (Fitzmaurice, 1995; Pepe & Anderson, 1994). Moreover, Fitzmaurice (1995) warned that the efficiency of an Independent GEE estimator that was associated with TDCs depended on the strength of the correlation between the responses; efficiency decreased drastically as the ignored correlation among responses increased. Although the use of the independent working correlation structure was recommended, in general, when time-dependent covariates were present, assuming independence between responses for subject *i* at different times could result in decreased efficiency of parameter estimates associated with that covariate (Fitzmaurice, 1995). In other words, the more significant the information being ignored, the greater the loss in efficiency.

Generalized Method of Moments

An alternate approach that could be taken in estimating parameters for marginal models of correlated data was to implement Generalized Method of Moments estimation (Hansen, 2007). The Generalized Method of Moments (GMM), like the Generalized Estimating Equations, is a method that accounts for correlation inherent in the data due to repeated measurements taken on the same subjects. Unlike GEE, GMM relies on the use

of moment conditions, expressions with zero expectation, rather than on the derivation of the likelihood or quasi-likelihood functions.

The process behind GMM involves minimizing a quadratic form, Q_N, over the parameters $\boldsymbol{\beta}$ (Hansen, 1982, 2007; Lai & Small, 2007):

$$Q_N(\boldsymbol{\beta}) = \boldsymbol{G_N}(\boldsymbol{\beta})^T \boldsymbol{W_N}\, \boldsymbol{G_N}(\boldsymbol{\beta}), \tag{6}$$

where $\boldsymbol{W_N}$ is a weight matrix. The vector $\boldsymbol{G_N}$ in Equation (6) is the average of moment conditions for all N subjects:

$$\boldsymbol{G_N} = \frac{1}{N}\sum_{i=1}^{N} g(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{\beta_0}),$$

where $\boldsymbol{Y_i}$ is the vector of responses for subject i, $\boldsymbol{X_i}$ is the vector of covariates for subject i, $\boldsymbol{\beta_0}$ is the vector of true parameters, and $g(\boldsymbol{Y_i}, \boldsymbol{X_i}, \boldsymbol{\beta_0})$ would be denoted $\boldsymbol{g_i}(\boldsymbol{Y_i}, \boldsymbol{X_i})$ for short.

When time-dependent covariates are present, the vector $\boldsymbol{g_i}(\boldsymbol{Y_i}, \boldsymbol{X_i})$ is composed of only the moment conditions that are considered "valid" for subject i, defined by Lai and Small (2007) as satisfying the expression:

$$E\,[\boldsymbol{g_i}(\boldsymbol{Y_i}, \boldsymbol{X_i})] = \boldsymbol{0}\,.$$

Lai and Small (2007) proposed using moment conditions that were products of derivative and residual terms at different times:

$$g_i = \frac{\partial p_{is}}{\partial \beta_j}(y_{it} - p_{it}),$$

where p_{is} is the mean response for subject i at time s (s = 1, ..., T), y_{it} is the response value for subject i at time t (t = 1, ..., T), p_{it} is the mean response for subject i at time t, β_j is the j^{th} covariate, i (i = 1, ..., N) denotes the subject, and $\boldsymbol{g_i}$ denotes the vector of valid moment conditions for subject i. Moment conditions are selected based on the type of time-

dependent covariates included in an analysis. For continuous data, p_{is} and p_{it} are replaced by the average response at time *s* and at time *t*, μ_{is} and μ_{it}, respectively. Further, it has been shown that the optimal choice for the weight matrix, $\boldsymbol{W_N}$, in Equation (6) is to use the inverse of the covariance matrix of the moment conditions (Hansen, 1982):

$$\boldsymbol{W_N} = \hat{\boldsymbol{V}}_{\boldsymbol{N}}^{-1} = Cov^{-1}\left(\boldsymbol{g_i}\right).$$

There have been several variations of the Generalized Method of Moments procedure, such as the Continuously Updating GMM, 2-Step GMM, and Iterative GMM (Chaussé, 2010; Hall, 2005; Hansen, 1982, 2007; Nielsen, 2005; Zivot, 2015). The difference in these types is determined by the choice of the weight matrix applied in the quadratic form that is minimized to obtain parameter estimates, and hence, the standard errors vary slightly. The quadratic forms minimized in the process behind the two most commonly used GMM types, Continuously Updating GMM (CUGMM) and 2-Step GMM (2SGMM), are:

$$QF_{CUGMM} : \boldsymbol{G_N}\left(\boldsymbol{\beta}\right)^T \boldsymbol{V_N}^{-1}\left(\boldsymbol{\beta}\right)\boldsymbol{G_N}\left(\boldsymbol{\beta}\right)$$

$$QF_{2SGMM} : \boldsymbol{G_N}\left(\boldsymbol{\beta}\right)^T \hat{\boldsymbol{V}}_{\boldsymbol{N}}^{-1}\left(\hat{\boldsymbol{\beta}}\right)\boldsymbol{G_N}\left(\boldsymbol{\beta}\right).$$

To obtain a 2SGMM estimator, an arbitrary initial weight matrix is selected, such as the identity matrix. This weight matrix is used to obtain the initial inefficient GMM estimator, $\hat{\boldsymbol{\beta}}_{initial}$. Using this inefficient estimator, an optimal weight matrix could be found, and this optimal weight matrix is used to obtain an efficient GMM estimator, $\hat{\boldsymbol{\beta}}_{efficient}$ (Hansen, 1982; Mátyás, 1999; Nielsen, 2005). Due to its dependence on the choice of the initial weight matrix, the 2SGMM estimator is not unique. Continuously Updating GMM (CUGMM) estimators, on the other hand, do not depend on the initial

weight matrix; rather, the weight matrix depends on the parameters (Nielsen, 2005; Zivot, 2015). The CUGMM estimation process simultaneously estimates the parameters, $\boldsymbol{\beta}$, and the weight matrix as a function of the parameters, $\boldsymbol{W}(\boldsymbol{\beta})$ (Mátyás, 1999; Zivot, 2015). Hence, with 2SGMM, $\hat{\boldsymbol{V}}_{\boldsymbol{N}}^{-1}\left(\hat{\boldsymbol{\beta}}\right)$ is fixed during the minimization of the quadratic form; whereas with CUGMM, the weighting matrix changes when $\boldsymbol{\beta}$ is changed in the minimization process (Chaussé, 2010; Mátyás, 1999; Nielsen, 2005; Zivot, 2015).

In obtaining the covariance matrix used to construct the quadratic form, Q_N, for obtaining GMM parameter estimates, Hansen (2007) additionally suggested the use of an iterative procedure in which an initial consistent GEE estimator $\hat{\boldsymbol{\beta}}^{(0)}$ is used to obtain $cov\{\boldsymbol{g}(\boldsymbol{Y_i}, x_{it}, \boldsymbol{\beta_0})\}^{-1}$, then estimating $\boldsymbol{\beta}_{\boldsymbol{GMM}}$ using $\hat{\boldsymbol{V}}_{\boldsymbol{N}}^{-1}$, yielding an estimator that is as asymptotically efficient as the traditional 2-Step GMM estimator and has consistent asymptotic variance (Hansen, 2007; Lai & Small, 2007), given by:

$$\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\,\boldsymbol{g}\left(\boldsymbol{Y_i}, x_{it}, \boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}\right)^{T} \times \left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{g}\left(\boldsymbol{Y}_i, x_{it}, \hat{\boldsymbol{\beta}}^{(0)}\right) g\left(Y_i, x_{it}, \hat{\beta}^{(0)}\right)^{T}\right)^{-1} \times \left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\,\boldsymbol{g}\left(\boldsymbol{Y_i}, x_{it}, \boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}\right)^{-1}$$

where $\frac{\partial\,\boldsymbol{g}\left(\boldsymbol{Y}_i, x_{it}, \boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}$ is evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{GMM}$ (Hansen, 2007). This research focused on the use of the 2-Step GMM.

In comparison to using the Independent GEE approach, the use of GMM estimation improved efficiency when time-dependent covariates were present. Results from the simulation study by Lai and Small (2007) showed that GMM estimators were more efficient than Independent GEE estimators when time-dependent covariates of types I or II were involved, and they were equally as efficient as Independent GEE estimators when a TDC of Type III was present. Moreover, in general, GMM estimators were

equally as efficient as GEE estimators when the working correlation structure was correctly specified, and they were asymptotically more efficient than GEE estimators when the working correlation structure was misspecified (Lai & Small, 2007). For these reasons, GMM estimation was a superior method when compared to Independent GEE when estimating parameters of a longitudinal study involving time-dependent covariates.

Statistics to Assess Model Goodness-of-Fit

For classical maximum likelihood estimation of independently observed data, some common statistics used in assessing model fit include the model deviance (Agresti, 1990; Dobson & Barnett, 2008; Pregibon, 1981), Akaike's Information Criterion, or AIC (Akaike, 1973, 1974), and Schwarz's Bayesian Information Criterion, or BIC (Schwarz, 1978). When using the Generalized Estimating Equations to estimate model parameters for correlated data, a modification of the AIC, known as the quasi-likelihood information criterion, or QIC (Pan, 2001a), and its adjustment, QICu, are used.

Model Deviance

To assess the fit of any generalized linear model, statisticians commonly use a goodness-of-fit measure known as the model deviance. Deviance gives a measure of the deviation of a specific model from the data. The likelihood evaluated using the data yields a model of "perfect fit" (Agresti, 1990). For discrete predictors, the model has as many parameters as observations. Deviance is a useful tool in assessing the goodness-of-fit of a model in that it could be used to compare a specific model to the full data, and it could also be used to compare two nested models.

When comparing a specific model to the data, deviance is given by:

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = 2\left[L(\boldsymbol{y};\boldsymbol{y}) - L(\hat{\boldsymbol{\mu}};\boldsymbol{y})\right],$$

where $L(\hat{\boldsymbol{\mu}}; \boldsymbol{y})$ is the maximum value of the log-likelihood under the given model, and $L(\boldsymbol{y}; \boldsymbol{y})$ is the maximum of the log-likelihood under the full data (Agresti, 1990). In other words, the deviance of a specific model, compared to the saturated model (i.e., the full data), is the difference in log-likelihoods under these two models. Because the log-likelihood is used, deviance could also be thought of as the log of the ratio of the likelihoods under the two models.

It is also known that:

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) \sim \chi^2_{(N-k)},$$

where N denotes the number of observations in the data and k denotes the number of parameters in the specified model (Agresti, 1990). The sampling distribution for the deviance results directly from the difference in log-likelihoods of two candidate models (Dobson & Barnett, 2008). For example, if $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator for the parameters $\boldsymbol{\beta_0}$ of a "true" model M_0, the difference in log-likelihoods of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta_0}$ could be written approximately:

$$l(\boldsymbol{\beta_0}; \boldsymbol{y}) - l(\hat{\boldsymbol{\beta}}; \boldsymbol{y}) = -\frac{1}{2}(\boldsymbol{\beta_0} - \hat{\boldsymbol{\beta}})^T \, \mathfrak{I}(\hat{\boldsymbol{\beta}}) \, (\boldsymbol{\beta_0} - \hat{\boldsymbol{\beta}}),$$

where $\mathfrak{I}$ is Fisher's information matrix.

For linear models, Fisher's information matrix is equivalent to:

$$\mathfrak{I} = \frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{X} \ ,$$

where $\boldsymbol{X}$ is the design matrix of the specified model (Dobson & Barnett, 2008).

For nonlinear models, Fisher's information matrix is equivalent to:

$$\mathfrak{I} = \frac{\partial^2 l}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_j}.$$

Then, the following statistic:

$$2\left[l\left(\hat{\boldsymbol{\beta}}; \boldsymbol{y}\right) - l\left(\boldsymbol{\beta}_0; \boldsymbol{y}\right)\right] = \left(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}\right)^T \mathfrak{I}\left(\hat{\boldsymbol{\beta}}\right)\left(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}\right)$$

is distributed as a chi-squared distribution $\chi^2_{(p-k)}$ with p = number of parameters in the "true" model, and k = number of parameters in the candidate model.

When comparing two nested models, models M_1 and M_2, where M_2 is nested within M_1, the deviance is given by:

$$D\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_2\right) - D\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_1\right) = 2\left[L\left(\hat{\boldsymbol{\mu}}_2; \boldsymbol{y}\right) - L\left(\hat{\boldsymbol{\mu}}_1; \boldsymbol{y}\right)\right] \sim \chi^2_{(p1-p2)},$$

where p_1 - p_2 denotes the difference in the number of parameters in the two models (Agresti, 1990). Since model M_2 is nested within M_1, the parameters of the vector $\boldsymbol{\beta_2}$ for M_2, is a subset of the parameter vector $\boldsymbol{\beta_1}$ for M_1.

No matter the models being compared--whether it be two nested models or a specific model compared to the saturated model--deviance is calculated as the difference in the log-likelihood of two models, or the likelihood ratio of two models (Agresti, 1990). It is the information not explained by the smaller model. It is similar to the Mean Squared Error (MSE) of a regression model, which gives an estimate of unexplained random error in the model. Moreover, model deviance is always distributed as a chi-squared distribution with degrees of freedom equal to the difference in the number of model parameters of the two models being compared.

Because the distribution of the model deviance is always known, hypothesis tests could be formed to assess model fit. The null hypothesis assumes the model fit for the

smaller model to be "sufficient." Smaller values of the chi-squared statistic suggest less deviation of overall goodness-of-fit of the candidate model from the "null" model (i.e., the data). In general, a small value of the deviance suggests that the smaller model is just as informative as the "null" model. The alternative, evidence to reject the null hypothesis, indicated by a large chi-squared test statistic, assumes poor fit for the smaller model.

Akaike's Information Criterion

To assess the fit of models derived using maximum likelihood estimation, researchers oftentimes rely on an information-based fit statistic known as Akaike's Information Criterion (AIC), proposed by Akaike (1973). Akaike's Information Criterion assumes no distribution; rather, it is a single value that is used descriptively to represent the amount of information lost from fitting a specific model to the data. It does not require the comparison of two nested models but rather compares a model to the actual data (Akaike, 1973).

Due to the asymptotic efficiency of maximum likelihood estimators under regularity conditions, Akaike (1974) suggested that the likelihood function tends to be very sensitive to small deviations of $\hat{\boldsymbol{\beta}}$ around $\boldsymbol{\beta_0}$. This property allows the "fit" of a model to be measured by the quantity:

$$S\left(g; f\left(\cdot|\hat{\boldsymbol{\beta}}\right)\right)=\int g\left(\boldsymbol{x}\right)\ln\left(\ f\left(\boldsymbol{x}|\hat{\boldsymbol{\beta}}\right)\right),$$

where the structure of a probability density function given by $f\left(\boldsymbol{x}|\hat{\boldsymbol{\beta}}\right)$ is compared to the structure of another pdf, $g(\boldsymbol{x})$. The probability density of $f\left(\boldsymbol{x}|\hat{\boldsymbol{\beta}}\right)$ could be used to define any model obtained using the classical maximum likelihood procedure with varying restrictions on $\hat{\boldsymbol{\beta}}$ (Akaike, 1974).

The result of these findings shed light on the importance of the role played by the functions

$$\frac{\partial}{\partial \boldsymbol{\beta}_j} \ln\left(f\left(\boldsymbol{x} | \hat{\boldsymbol{\beta}}\right)\right)$$

in Akaike's definition of "information lost" due to deviations of $\hat{\boldsymbol{\beta}}$ from $\boldsymbol{\beta_0}$ (Akaike, 1973). More specifically, it was understood that the difference in the values of the maximum of the log-likelihoods:

$$2\left(\sum_{i=1}^{N} \ln\left(f\left(x_i | \boldsymbol{\beta_0}\right)\right) - \sum_{i=1}^{N} \ln\left(f\left(x_i | \hat{\boldsymbol{\beta}}\right)\right)\right),$$

represents a factor of the estimate of the discrepancy between $\boldsymbol{\beta_0}$ and $\hat{\boldsymbol{\beta}}$ (Akaike, 1973, 1974). Akaike's Information Criterion (AIC) is defined as:

$$AIC = -2\sum_{i=1}^{N} \ln\left(f_i\left(y_i; \hat{\beta}_i\right)\right) + 2k$$

$$= -2\,L\left(\hat{\boldsymbol{\beta}}; \boldsymbol{y}\right) + 2k\,, \tag{7}$$

where f_i is a probability density function determined by the distribution of the response variable y_i, $L\left(\hat{\beta}; y\right)$ is the maximum of the log-likelihood of the specified model, and k is the number of parameters in the model, not including interaction terms. The penalty term $2k$ in Equation (7) is added to this quantity to account for the bias introduced by estimating $\boldsymbol{\beta_0}$ by $\hat{\boldsymbol{\beta}}$ (Akaike, 1974).

Akaike's Information Criterion (AIC) is an information-based theoretical criterion that could be used to compare nested models with varying numbers of parameters (Akaike, 1974), and models with more parameters are penalized to a greater extent. Thus, AIC could be thought of as a fit statistic that selects parsimonious models. The larger the

value of the AIC, the farther the model deviates from the observed data. Better fitting models are associated with having smaller AIC values (Akaike, 1973). In the sense of multiple regression models, the value of AIC for models with different combinations of predictors could be used to select which model is “ideal” (Akaike, 1974).

Ever since the AIC was introduced as a method of model selection, it has become very popularly used in various disciplines, especially when model parameters have been estimated using a likelihood-based method. An advantage of using AIC for model selection is that it could be applied to both nested and non-nested models (Sayyareh, Obeidi, & Bar-Hen, 2011); however, it performs best in large-sample situations in which the true model is included among the set of candidate models (Kuha, 2004). Additionally, as AIC represents the amount of information lost from approximating the full data using a model, it could not be considered a hypothesis test, and thus, the results of the model selection process using AIC could be inconclusive in determining the best model (Sayyareh et al., 2011). Moreover, there is debate about its liberalness in commonly selecting complex models beyond necessity (Sayyareh et al., 2011).

A major drawback of the AIC is that its bias could be substantial in very small samples, and the use of an adjusted AIC, such as the corrected AIC (AICc), is recommended in its stead to correct for finite-sample bias (Burnham & Anderson, 1998; Hurvich & Tsai, 1989; Kuha, 2004; Sugiura, 1978). In fact, when the sample size is small, AIC tends to favor models that are far too parsimonious, even more parsimonious than those selected by the small-sample corrected AIC (Ward, 2008); thus, the information criterion that includes an adjustment should be used over AIC when small samples are under study (Burnham & Anderson, 1998; Hurvich & Tsai, 1989; Kuha,

2004; Posada & Buckley, 2004; Sugiura, 1978). On the other hand, AIC tends to select more complex models when large samples are considered (Kuha, 2004; Posada & Buckley, 2004; Ward, 2008), and therefore, it is not considered a consistent method of model selection (Kuha, 2004; Ward, 2008).

Corrected Akaike's Information Criterion

When building and comparing candidate models for small-sample data, a small-sample bias correction is available for Akaike's Information Criterion (Hurvich & Tsai, 1989, 1995). This small-sample version of AIC, known as corrected AIC, or AICc, is:

$$AIC_C = AIC + \frac{2(k+1)(k+2)}{n-k-2},$$

where *k* is the number of parameters in the estimated model, *n* is the number of observations in the sample, and AIC is Akaike's Information Criterion characterized by Equation (7) in the preceding section. It has been shown that, as the sample size grows larger, AICc converges to AIC (Burnham & Anderson, 2004; Hurvich & Tsai, 1995):

$$\lim_{n\to\infty} \text{AICc} = \text{AIC}.$$

Due to the common and inappropriate use of AIC over AICc in situations involving small sample sizes, some researchers have encouraged the regular use of AICc in lieu of AIC, arguing that AICc converges to AIC with increasing sample size (Anderson & Burnham, 2002; Burnham & Anderson, 2004).

Bayesian Information Criterion

Another commonly used information criterion that measures the goodness-of-fit of a model is the Bayesian Information Criterion (BIC), proposed by Schwarz (1978):

$$BIC = -2\sum_{i=1}^{N} \ln\left(f_i\left(y_i;\hat{\beta}_i\right)\right) + k\ln(n)$$

$$= -2L\left(\hat{\boldsymbol{\beta}};\boldsymbol{y}\right) + k\ln(n),$$

where f_i is a probability density function determined by the distribution of the response variable $\boldsymbol{y}_i$, $L\left(\hat{\boldsymbol{\beta}};\boldsymbol{y}\right)$ is the maximum of the likelihood derived from the data, k is the number of predictors in the model, and n is the number of observations in the data.

Bayesian Information Criterion (BIC), also known as the Schwarz Information Criterion (Cavanaugh & Neath, 1999), is very similar to AIC in that it uses the maximum value of the log-likelihood of the model, and it penalizes for the number of parameters included in the model. The penalty for BIC (i.e., the $k \cdot ln(n)$ term) is larger than that of the AIC because BIC penalizes a model for the number of parameters it includes, as well as the number of observations in the data. The larger the number of observations in the data, the larger the value of $ln(n)$ becomes. For this reason, BIC tends to favor more parsimonious models than AIC.

Although both AIC and BIC favored parsimonious models, BIC tends to lean more towards lower-dimensional models with a smaller sample size because the process behind BIC assigns more penalty to models with larger n (Schwarz, 1978). For data with smaller sample sizes, BIC assigns more weight to complex models compared to AIC, and it assigns less weight to simpler models as n increases (Kuha, 2004; Posada & Buckley, 2004; Ward, 2008). In general, the selection of models using AIC and BIC differs vastly for large numbers of observations (Cavanaugh & Neath, 1999; Schwarz, 1978).

The results of simulation studies by Wang and Liu (2006) showed that BIC tended to be a better model selection method than AIC when comparing nested models

(Wang & Liu, 2006). However, as was the case with AIC, BIC could not be considered a form of hypothesis testing, as it is not distribution-based and is used as information to represent information lost from fitting a model to the data (Sayyareh et al., 2011).

In further comparing BIC with AIC as a method of model selection, BIC assumes that the true model exists among the set of candidate models and attempts to identify it, whereas AIC assumes that there is no way of identifying a true model among the candidates and selects the one that most adequately fits the data (Kuha, 2004). In other words, AIC is used to select a model that best approximates reality, whereas the purpose of BIC is to identify the true model among the set of candidate models (Posada & Buckley, 2004). However, when "incorrect" models are selected, BIC tends to favor models that are too parsimonious (Kuha, 2004).

Bayesian Information Criterion (BIC) is constructed using the assumptions related to Bayesian prior distributions, which may not necessarily correspond to reality (Weakliem, 1999). As a result, when the assumed prior distribution does not reflect reality, BIC may not necessarily represent the actual discrepancy of information between the model and the data. Beside this possible bias, however, BIC is a consistent method of model selection, as its penalty term, $k \cdot ln(n)$, is an increasing function of n (Kuha, 2004; Ward, 2008), and the results of using BIC for model selection is not affected significantly regardless of the assumed Bayesian prior information (Weakliem, 1999).

Quasi-Likelihood Information Criterion

When assessing the fit of models with parameters estimated using maximum likelihood estimation procedures, either the model deviance or information-based goodness-of-fit statistics, AIC and BIC, could be used. These statistics required the use of

the full likelihood of the response. When using the Generalized Estimating Equations approach, we used a quasi-likelihood instead of the full likelihood. Analogously, in assessing the fit of models with parameters estimated using GEE, the quasi-likelihood information criterion (QIC) was commonly used (Pan, 2001a).

The quasi-likelihood function is defined by the equation:

$$Q(p;y)=\int_{y}^{p}\frac{y-m}{\sigma^2 Var(m)}dm,$$

where p is the average response, y is the observed response, σ^2 is the error variation, and m is an arbitrary observation (McCullagh & Nelder, 1989; Wedderburn, 1974). For a binary response, the quasi-likelihood function is given by:

$$Q(p;y)=y\cdot\ln\left(\frac{p}{1-p}\right)+\ln(1-p),$$

where p is still the average response, and y is the observed response (McCullagh & Nelder, 1989). Derivation of the quasi-likelihood function only required knowledge of the mean and mean-variance relationship of the response, rather than full knowledge of the response distribution, which was required in deriving the likelihood function. Consequently, the quasi-likelihood was based on the first two moments of the response, rather than on the marginal distribution of the response, as was the case with the full likelihood function. Other properties of the quasi-likelihood could be found in Wedderburn (1974) and McCullagh and Nelder (1989).

A measure of separation between two models could be given by the Kullback-Leibler information (Kullback & Leibler, 1951):

$$\Delta(\boldsymbol{\beta_1},\boldsymbol{\beta_0})=E_{M_0}\left[-2L_1(\boldsymbol{\beta_1},\boldsymbol{Y})\right],$$

where M_0 denotes the "true" model, M_1 denotes a candidate model nested within M_0, L_1 represents the log-likelihood function of the candidate model, M_1, and $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_0}$ are the estimates obtained under the candidate and true models, respectively. From a set of all possible candidate models, the model with the smallest value of $\Delta(\boldsymbol{\beta_1}, \boldsymbol{\beta_0})$ is selected.

Bridging this measure of distance to establish some estimate for the discrepancy of competing models derived using the quasi-likelihood rather than the full likelihood, Pan (2001a) replaced the likelihood function with the quasi-likelihood in the Kullback-Leibler equation:

$$\Delta\left(\boldsymbol{\beta_1}, \boldsymbol{\beta_0}, \boldsymbol{I}\right) = E_{M_0}\left[-2Q\left(\boldsymbol{\beta_1}, \boldsymbol{I}, \boldsymbol{Y}\right)\right],$$

where $\boldsymbol{I}$ is the independent working correlation structure, Q is the quasi-likelihood under the working independence model, and $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_0}$ are the estimates obtained under the candidate and true models, respectively. Pan (2001a) suggested approximating $E_{M_0} \approx \left[\Delta\left(\hat{\boldsymbol{\beta}}_{\mathbf{1}}, \boldsymbol{\beta_0}, \boldsymbol{I}\right)\right]$ using:

$$E_{M_0} \approx \left[\Delta\left(\hat{\beta}_1, \beta_0, I\right)\right] - 2E_{M_0}\left[Q\left(\hat{\beta}_1, I, Y\right)\right] + 2E_{M_0}\left[\left(\hat{\beta}_1 - \beta_0\right)^T S\left(\hat{\beta}_1, I, Y\right)\right] + 2trace\left(\Omega_I, \hat{V}_r\right), \quad (8)$$

where $\boldsymbol{\Omega_I} = \sum_{i=1}^{N} \boldsymbol{D_i}^T \boldsymbol{V_i}\, \boldsymbol{D_i}$, $D_i = \dfrac{\partial \boldsymbol{\mu_i}\left(\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}^T}$ is the matrix of partial derivatives of the mean for subject i with respect to the parameter vector, $\boldsymbol{V_i}$ is a working covariance matrix of $\boldsymbol{Y_i}$, and $\hat{V}_r$ is the consistent robust or sandwich covariance estimator proposed by Liang and Zeger (1986).

Ignoring the second term in Equation (8), which is difficult to estimate, gives the goodness-of-fit statistic known as the quasi-likelihood information criterion (QIC) proposed by Pan (2001a):

$$QIC = -2Q\left(\hat{\boldsymbol{\beta}}\left(\boldsymbol{R}_i\left(\boldsymbol{\alpha}\right)\right); \boldsymbol{I}, \boldsymbol{Y}\right) + 2\,trace\left(\boldsymbol{\Omega}_{\boldsymbol{I}}\hat{\boldsymbol{V}}_r\right) ,$$

where Q is the quasi-likelihood of the response under the independence model, $\hat{\boldsymbol{\beta}}$ is the vector of parameter estimates for the candidate model, and the efficiency of parameter estimates depends heavily on the selected working correlation structure, $\boldsymbol{R_i(\alpha)}$. In his simulation studies, Pan (2001a) showed that ignoring the second term in Equation (8) did not influence the results of QIC drastically. It has been shown, however, that the working independence quasi-likelihood based model showed best performance in terms of model efficiency (Pan, 2001a).

The quasi-likelihood information criterion (QIC) could be used to select a working correlation structure in GEE. The QIC is calculated using $\hat{\boldsymbol{\beta}}$ based on various working correlation structures, and the working correlation structure with the smallest associated QIC is selected (Pan, 2001a). Using this information criterion in this way, a small value for the QIC suggests that the specified working correlation is very close to the true correlation structure of the data or is the optimal choice for the data.

An Alternative to the Quasi-likelihood Information Criterion

An alternative statistic to the QIC is also available. Pan (2001a) showed that when all modeling specifications in GEE were correct, $\hat{\boldsymbol{\Omega}}_{\boldsymbol{I}}^{-1}$ was asymptotically equivalent to $\hat{\boldsymbol{V}}_r$, and $trace\left(\hat{\boldsymbol{\Omega}}_{\boldsymbol{I}}^{-1}\hat{\boldsymbol{V}}_r\right) = trace\ (\boldsymbol{I}) = k$, then the QIC is equivalent to the AIC (Hosmer et

al., 2013; Pan, 2001a)--with the quasi-likelihood in place of the likelihood--and could be estimated using:

$$QICu \equiv -2Q\left(\hat{\boldsymbol{\beta}}\left(\boldsymbol{R}_i\left(\boldsymbol{\alpha}\right)\right); \boldsymbol{I}, \boldsymbol{Y}\right) + 2k \,,$$

where, again, *k* is the number of parameters in the specified model (Pan, 2001a).

The uses and applications of QIC and QICu varied in the literature: QIC was used to select a working correlation structure that best fit the data, whereas QICu was used to determine which set of predictors best explained the response under investigation (Hilbe, 2009; Hosmer et al., 2013). To select a model, the QICu was calculated for all candidate models, and the model with the smallest QICu was selected.

The alternative quasi-likelihood information criterion (QICu) could be used for variable selection, but it could not be used to select the working correlation structure because the penalty term of the QICu ignores the correlation structure (Hosmer et al, 2013; Pan, 2001a). On the contrary, the penalty term of the QIC includes the correlation structure; hence, the QIC may have potential use in both selection of the most appropriate working correlation structure, as well as variable selection (Hosmer et al., 2013).

Results of the GEE analysis for the National Longitudinal Survey data conducted by Cui (2007) showed that both the QIC and QICu selected the same model, the full model, in two different examples; this would not always be the case. When different models were selected by these two measures, it was advised that the QIC be used for model selection, as the QICu was only an approximation to the QIC (Cui, 2007; Hilbe, 2009; Hosmer et al., 2013).

Goodness-of-Fit of Models Estimated Using Generalized Method of Moments

The literature was sparse in its discussions of the assessment of fit of models with parameters estimated using the Generalized Method of Moments. Existing measures of fit--including deviance, AIC, BIC, QIC, and QICu--were not appropriate statistics in assessing the fit of models in which parameter estimates were obtained using GMM because these statistics were derived using either the full likelihood or the quasi-likelihood of the parameters, while GMM does not employ the use of any likelihood function.

The majority of discussions involving GMM and model fit revolved around the idea of overidentification of models due to misspecification (Andrews, 1999; Hall, 1999; Hansen, 1982; Hansen, Heaton, & Yaron, 1996; Imbens, 1997; Newey, 1985). Overidentification results when more than the necessary number of moment conditions are used in the estimation of model parameters; this suggests that the model is misspecified and that an alternative model should be considered (Hall, 1999; Newey, 1985). The identifying and overidentifying restrictions decompose the population moment conditions into two orthogonal pieces: the identifying restrictions are the components of the population moment conditions that are used in the estimation of model parameters; the overidentifying restrictions are the remaining components (Hall, 1999; Hansen, 1982; Hansen et al., 1996). This decomposition comes from minimizing the quadratic form, $Q_N(\boldsymbol{\beta})$, to obtain GMM parameter estimates. Misspecification could lead to modeling issues, as it could cause parameter estimates to be inconsistent and invalidate common practices behind inferential procedures (White, 1982).

However, overidentification did not directly address the primary interest of this research. While the concern with overidentification was with selections of specific moment conditions used in the Generalized Method of Moments estimation, the main focus of this research was on the selection of predictors for modeling longitudinal data with time-dependent covariates.

Distributions of Differences in Minimands of Quadratic Forms

Useful results about the distribution of the minimum of the GMM quadratic form, $Q_N\left(\hat{\boldsymbol{\beta}}\right)$, were presented by Hansen et al. (1996). It has been shown that the minimand of the quadratic form multiplied by the number of subjects, *N*, is distributed as a chi-squared with degrees of freedom equal to the difference between the number of moment conditions in the population and the number of estimated parameters (Hansen et al., 1996):

$$\min\left(Q_N\left(\hat{\boldsymbol{\beta}}\right)\right) \sim \chi^2_{\,(\mathrm{m}\,-\,\mathrm{k})}, \tag{9}$$

where *m* denotes the number of moment conditions used in the estimation process, and *k* denotes the number of parameters in the model.

Results presented in the form of Equation (9) allowed researchers to test hypotheses about the quadratic form, which gave insight as to whether or not "sufficient" moment conditions were used in the estimation process. However, this assessed the appropriate inclusion--or exclusion--of a set of moment conditions rather than assessing the overall fit of the model. This addressed overidentification but has not been implemented for selection of predictor variables.

The results from Newey (1985) and Hall (1999) pointed in a similar direction. Hypothesis tests for the overidentifying moment conditions that resulted in misspecification were discussed (Hall, 1999), using the chi-squared test from Hansen (1982) and Hansen et al. (1996). The low power and inconsistency of these misspecification tests were presented in great mathematical detail (Newey, 1985; White 1982); however, very little was discussed about the overall goodness-of-fit of the model. Moreover, there was no discussion in the context of time-dependent covariates or any discussion involving the selection of predictor variables.

Hypothesis Tests for Selection of Moment Conditions

Discussions involving a Goodness-of-fit (GOF) like statistic presented by Lai and Small (2007) introduced hypothesis tests for selection of appropriate moment conditions in order to evaluate time-dependent covariate type. The statistic used in these hypothesis tests relied on the difference in minimands of quadratic forms under the two models fitted using GMM. Let M_1 and M_2 denote two candidate models and let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ be the vector of parameter estimates obtained under these models, respectively. Define the statistic C_N as:

$$C_N = N\left[\min\left\{QF\left(\hat{\boldsymbol{\beta}}_1\right)\right\} - \min\left\{QF\left(\hat{\boldsymbol{\beta}}_2\right)\right\}\right] \sim \chi^2_{\text{r-q}}, \qquad (10)$$

where $QF\left(\hat{\boldsymbol{\beta}}_1\right)$ and $QF\left(\hat{\boldsymbol{\beta}}_2\right)$ are minimands of the quadratic forms of two candidate models, and $r - q$ is the difference in the number of moment conditions used in the GMM estimation of these models. Tests using the statistic in Equation (10) were used to select appropriate moment conditions, equivalent to testing whether or not the specification of the type of time-dependent covariate was correct (Lai & Small, 2007). The null

hypothesis assumed that a subset of the population of moment conditions was sufficient in the GMM estimation process; rejecting this hypothesis suggested that an incorrect subset of moment conditions had been selected. The quadratic form used in the GMM process depends on different choices of moment conditions; hence, if an inappropriate set of moment conditions is selected, inconsistent GMM parameter estimates may result (Lai & Small, 2007). However, this process selects the type of time-dependent covariate rather than predictors to include in a model.

The main goal of this research was to establish a fit statistic--an information-based criterion, much like AIC or BIC--that has no distribution but is represented by a scalar to compare nested models that are fit using the Generalized Method of Moments. Specifically, interest was geared toward establishing a statistic for the fit of models in which parameter estimates were obtained using GMM using data with at least one time-dependent covariate, as well as on the selection of predictors when comparing models with time-dependent covariates.

Rationale for Research

Currently, there have been very few methods for assessing the goodness-of-fit of models in which parameters were estimated using the Generalized Method of Moments in the presence of time-dependent covariates. One approach was to use the minimum of the quadratic form to assess the overall "goodness" of the model; the smaller the minimum of the quadratic form, the better the model fit. Distributional results from Hansen et al (1996), as well as the discussion on non-nested models by Hall (1999), could be used to form hypothesis tests for overidentifying restrictions of moment conditions. However, these methods test whether or not an estimate $\hat{\beta}$ formed from a set of moment conditions

deviates from $\boldsymbol{\beta_0}$; they did not test overall model fit or select appropriate predictors that were time-dependent covariates for the model.

Additionally, there was no published work in the body of literature that suggested any of these methods were employed in the presence of time-dependent covariates. There was a need for an information-based fit statistic to assess the overall goodness-of-fit of models in which parameters of a marginal model for longitudinal data with time-dependent covariates were estimated using GMM. Without such a statistic, it would be difficult to make comparisons between models with different time-varying predictor variables.

CHAPTER III

METHODOLOGY

Currently, there has been no information-based method for assessing the goodness-of-fit of models in which parameters were estimated using the Generalized Method of Moments. Additionally, there was no published work in the body of literature that suggested any existing method was employed in the presence of time-dependent covariates. There was the need to establish a statistic that was based on moment conditions rather than on the specification of the full likelihood or quasi-likelihood functions. Further, there was a need for a more information-based fit statistic to assess the overall goodness-of-fit of models, as well as the relative fit of nested models, in which parameters were estimated using GMM in the presence of time-dependent covariates. Without such a statistic, it would be difficult to assess whether or not a model and its parameter estimates are appropriate.

Chapter III outlines the process for obtaining two statistics that could be used in conjunction with each other to assess the overall fit of models constructed using the Generalized Method of Moments in the presence of time-dependent covariates. These methods relied neither on the likelihood, nor the quasi-likelihood, but rather on the use of moment conditions, as used in the GMM process.

Research Questions

This chapter outlines the methodology for answering the following research questions:

Q1 How can information associated with the fit of model parameters estimated using the Generalized Method of Moments be expressed or measured?

Q2 What is the detection proportion of the model selection process of such measures in their ability to detect poor fit of underfit models?

Q3 What are the non-detection proportions of such measures in suggesting poor fit for appropriate models?

Moment-Based Goodness-of-Fit Statistics

In this section, two measures were proposed that could be used in conjunction with each other to assess the goodness-of-fit of a model in which parameter estimates were obtained using the Generalized Method of Moments when time-dependent covariates were in the data. These methods relied on neither the likelihood function, nor the quasi-likelihood function; instead, they relied on the use of moment conditions and an established measure of distance. The first statistic utilized the minimum of the quadratic form used in the GMM estimation process (Hansen, 1982, 2007; Lai & Small, 2007), much like the chi-squared statistic used in Hansen et al. (1996). The second statistic was a variation of the Kullback-Leibler divergence principle (Kullback & Leibler, 1951) and was an information criterion, similar in use to the familiar AIC and BIC statistics.

Minimum of the Generalized Method of Moments Quadratic Form

One common approach used to assess the overidentification of models fit using GMM estimation in the literature was to examine the minimum of the quadratic form used in the GMM process (Hall, 1999; Hansen et al., 1996), as expressed in Equation (9).

Large values of the minimand of the quadratic form, which is the value that minimizes the quadratic form, suggested overidentification (Hall, 1999; Hansen et al., 1996). As mentioned in the section "Distributions of Differences in Minimands of Quadratic Forms," the statistic in Equation (9) was not used to assess the overall model fit but to determine whether a sufficient set of moment conditions was used in the estimation process.

Let M_1 and M_2 denote two candidate models fit using GMM, and M_2 is nested within M_1, and let $\hat{\boldsymbol{\beta}}_1$ be the parameters estimates obtained under M_1 and $\hat{\beta}_2$ be the parameters estimates obtained under M_2. Based on the findings presented in Hansen et al. (1996), the goodness-of-fit of two nested GMM models, M_1 and M_2, could be compared using the statistic:

$$C\left(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2\right) = \left(\min\left\{ QF\left(\hat{\boldsymbol{\beta}}_1\right)\right\} - \min\left\{ QF\left(\hat{\boldsymbol{\beta}}_2\right)\right\}\right) \sim \chi^2_{(p_1 - p_2)}, \tag{11}$$

where QF($\hat{\boldsymbol{\beta}}_1$) and QF($\hat{\boldsymbol{\beta}}_2$) are the quadratic forms of the two nested models, and $p_1 - p_2$ denotes the difference in the number of parameters between the two candidate models. In other words, the difference between the minimum of the GMM quadratic forms of two competing models follows a chi-squared distribution with degrees of freedom equal to the difference between the numbers of parameters in the two models (Hansen et al., 1996).

The statistic in Equation (11) could be used to test the hypothesis:

H_0: The candidate model with fewer parameters is sufficient in explaining the information in the response.

H_A: The candidate model with fewer parameters is *not* sufficient in explaining the information in the response.

If the true model, M_0, was known and fully specified using the full data, the statistic $C(\hat{\beta}_1, \hat{\beta}_2)$ could be used to assess whether an alternate model with fewer parameters is adequate in explaining the variability present in the response, due to trivial deviation from the full model (Hall, 1999; Hansen et al., 1996).

Although the chi-squared statistic has been popularly used to compare candidate models, there was no discussion of its use when model parameters were estimated using GMM when time-dependent covariates were present in the data. As models and parameter estimates varied based on the selection of different sets of moment conditions, this difference impacted the result of the chi-squared test for model selection and selection of predictors. Therefore, this research considered the use of this statistic in the presence of time-dependent covariates for the first time.

Kullback-Leibler Information Criterion

The Kullback-Leibler divergence principle is an information-based measure of distance, and it could be used as a measure to assess the discrepancy of a candidate model from the "null" model, which entails the full the data (Csiszár, 1975; Kullback & Leibler, 1951; White, 1982). Although the Kullback-Leibler divergence principle, also known as "*I*-divergence" from a purer mathematical standpoint, was not a formal metric of distance, the properties of probability distributions common to those of Euclidean geometry allowed for this information criterion to be used as a measure of distance (Csiszár, 1975). As an alternative to the more common optimally weighted GMM estimation procedure (Hansen, 1982), Kitamura and Stutzer (1997) proposed an estimator that was based on the minimization of the Kullback-Leibler information divergence for modeling dependent data:

$$\begin{aligned}\left(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}\right) &= \arg\max_{\boldsymbol{\beta}}\min_{\boldsymbol{\gamma}}\left\{\hat{Q}_T\left(\boldsymbol{\beta},\boldsymbol{\gamma}\right)\right\} \\ &= \arg\max_{\boldsymbol{\beta}}\min_{\boldsymbol{\gamma}}\left[\frac{1}{T}\sum_{t=1}^{T}\exp\left(\boldsymbol{\gamma}^T\ \hat{f}\left(t,\boldsymbol{\beta}\right)\right)\right],\end{aligned} \tag{12}$$

where f denotes a probability distribution function (PDF) characterized by observations $\boldsymbol{t}$, $\boldsymbol{\beta}$ is the vector of parameters of a model fit using GMM, $\boldsymbol{\gamma}$ is a vector of unknown parameters, T is the total number of repeated observations per subject, and Q_T is some known function of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (Kitamura & Stutzer, 1997). Equation (12) is used to simultaneously estimate the vector of model parameters, $\boldsymbol{\beta}$ and the vector of unknown parameters, $\boldsymbol{\gamma}$. In this research study, Equation (12) was not used to obtain $\boldsymbol{\beta}$; the parameters were obtained using the more common 2-Step Generalized Method of Moments (2SGMM), as used in the simulation study by Lai and Small (2007). Thus, the portion of Equation (12) involving maximizing over $\boldsymbol{\beta}$ was ignored in this research study, and the portion of Equation (12) involving $\boldsymbol{\gamma}$ was used to obtain the Kullback-Leibler divergence-based information criterion, which was called the Kullback-Leibler Information Criterion (KLIC).

The KLIC statistic could be used as a measure of "information lost" because other information criteria, such as the AIC, were derived from the general idea underlying the use of Kullback-Leibler divergence as a measure of information (Akaike, 1973, 1974; Kuha, 2004; Rodríguez, 2005; Sayyareh et al., 2011). In this research, a variation of Equation (12) was used to obtain a measure of information lost--the KLIC--rather than as a process to obtain parameter estimates, $\hat{\boldsymbol{\beta}}$, as was its use in Kitamura and Stutzer (1997).

Rather than using a probability distribution function (PDF), f, as in Equation (12), this study employed the use of moment conditions. From a fundamental standpoint,

generalized moments could be computed fairly easily, revealing important aspects of the underlying probability distribution: the first moment is the population mean, the second moment is the population variance, the third moment the skewness, and the fourth kurtosis (Zsohar, 2012). Moments contain information about the location, scale, and shape of the distribution, without a full specification of the underlying distribution of the data under investigation. Moment conditions could be used in lieu of a PDF because they share similar properties: both are functions of the parameters, $\boldsymbol{\beta}$, and could be estimated from the data (Csiszár, 1975; Hall, 1999; Hansen et al., 1996; Kitamura & Stutzer, 1997). Once the moments are obtained, sample statistics are associated with their population counterpart, such as the relation between the sample mean and the population expected value, and the sample moments are used as the foundation of the parameters to be estimated (Stigler, 2008; Zsohar, 2012).

The theoretical definition of the Kullback-Leibler Information Criterion is given by the equation:

$$D = -\ln\left(E\left[\exp\left(\gamma^T \, \boldsymbol{g}_i\left(\boldsymbol{\beta}\right)\right)\right]\right), \tag{13}$$

where $\boldsymbol{g}_i(\boldsymbol{\beta})$ is the vector of valid moment conditions for subject *i* used in the GMM process to obtain the quadratic form, and $\boldsymbol{\gamma}$ is a vector of unknown parameters (Altonji & Segal, 1996; Kitamura & Stutzer, 1997). For a sample, the statistic in Equation (13) could be estimated using:

$$\hat{D} = -\ln\left(\frac{1}{N}\sum_{i=1}^{N}\exp\left(\hat{\gamma}^T \, \hat{\boldsymbol{g}}_i\left(\hat{\boldsymbol{\beta}}\right)\right)\right),$$

where *N* denotes the total number of subjects, $\boldsymbol{g}_i$ is the vector of estimates of moment conditions using $\hat{\boldsymbol{\beta}}$ from 2SGMM, and $\hat{\gamma}$ is the vector of unknown parameters estimated

by the minimization of Equation (12) using $\hat{\boldsymbol{\beta}}$ from 2SGMM. A more computational form of the Kullback-Leibler Information Criterion, which was used in this research, was given by (Kitamura & Stutzer, 1997):

$$KLIC = \min_{\gamma} \left(\frac{1}{N} \sum_{i=1}^{N} \exp\left(\hat{\gamma}^T \, \hat{g}_i \left(\hat{\beta}\right)\right) \right). \tag{14}$$

As the natural log in Equation (13) is a monotonic function of the parameters of interest, its omission in Equation (14) did not alter the conclusions deduced from the use of the KLIC as an information criterion.

The process to obtain the KLIC was as follows: using the 2-Step Generalized Method of Moments (2SGMM; Lai & Small, 2007), parameter estimates, $\hat{\boldsymbol{\beta}}$, could be obtained, as well as the vector of valid moment conditions for each subject, $\boldsymbol{g_i}$, that was used to obtain those 2SGMM parameter estimates, $\hat{\boldsymbol{\beta}}$. Then, the vector of unknown parameters, $\boldsymbol{\gamma}$, which was used to minimize Equation (12), was obtained. These items were then placed into Equation (14), averaged across all *N* subjects, and minimized with respect to $\boldsymbol{\gamma}$. This minimization process yielded an estimate of the Kullback-Leibler Information Criterion in Equation (14).

The Kullback-Leibler Information Criterion is similar to other measures of goodness-of-fit, such as AIC and QIC, in that it is represented by a single number and follows no known distribution; therefore, no hypothesis test can be formed using the Kullback-Leibler information based measure of distance as a basis for inferential conclusions. Rather, the statistic could be used to compare nested models; the candidate model with the smallest value of the Kullback-Leibler Information Criterion was selected.

Real Data

In order to answer the research questions posed in this study, both real-world data and simulated data were used, and 2-Step Generalized Method of Moments (2SGMM) was used to estimate parameters for all models. All management, simulation, and analyses of both the real data and the simulated data were conducted using the software environment R version 3.4.2.

Real Data: A Study on the Health of Filipino Children

The data came from a longitudinal study conducted by members of the International Food Policy Research Institute in 1984-1985 on the health of Filipino children aged 1-10 years from the Bukidnon region of the island of Mindanao (Bhargava, 1994; Bouis & Haddad, 1990; Lai & Small, 2007). Four nutritional surveys were administered at 4-month intervals, measuring age, gender, height, weight, food consumed in the previous 24 hours, whether or not the child suffered from various illnesses in the previous 2 weeks, as well as the duration of the sickness (in days) from a total of 448 households. To minimize the association of errors of these measurements from children within the same household, the data from only the youngest child were maintained, and only those individuals with complete data at all time points were considered. This resulted in balanced longitudinal data with 3 observations each from 370 unique children. Bouis and Haddad (1990) and Bhargava (1994) offer more details about the data.

The Models for the Filipino Child Mortality Data

In order to construct a logistic regression model and a multiple linear regression model to predict a child's morbidity, new variables were first calculated from the data.

The focus of this research was on logistic regression but a continuous response was also considered for the sake of making comparisons to existing publications. The original data included a binary variable for morbidity, so no transformation was necessary to obtain the binary response. Following Bhargava (1994) and Lai and Small (2007), the dependent variable for the multiple linear regression model was the transformed morbidity outcome, which was a variable of the form:

$$y_{it} = \ln\left(\frac{t_{before} + 0.5}{14.5 - t_{before}}\right), \tag{15}$$

where t_{before} was the days over the previous 2 weeks before time t the child was sick. The transformation in Equation (15) yielded a continuous response for the sake of multiple linear regression. Rather than using height and weight as individual predictors of morbidity, which could be correlated for children between 1 and 10 years of age, a body mass index (BMI), the ratio of weight (in kilograms) and squared height (in meters), were calculated for each child. Gender, age, and the indicators for survey round were also included to model morbidity, but no transformation were necessary for these variables.

The equation for the models to predict morbidity 4 months in the future was,

$$\eta_{it} = \beta_0 + \beta_1 x_{it,Age} + \beta_2 x_{it,Gender} + \beta_3 x_{it,BMI} + \beta_4 x_{it,Round2} + \beta_5 x_{it,Round3} + \varepsilon_{it}, \tag{16}$$

where $x_{it,Age}$ was the covariate for age of subject i at time t, $x_{it,Gender}$ was the covariate for gender of subject i at time t, $x_{it,BMI}$ was the covariate for BMI of subject i at time t, $x_{it,Round2}$ and $x_{it,Round3}$ were the indicator variables for survey rounds 2 and 3, respectively, of subject i at time t, and ε_{it} was the error term. As logistic and linear regression models were considered, the systematic component $\eta_{it} = \mu_{it}$ was defined differently for each case: g was the identity link function for the linear regression model, where μ_{it} was the mean

response for subject i at time t, and g was the logit-link function for the logistic regression model, where $\mu_{it} = p_{it}$ was the probability of success for subject i at time t.

Of the five predictors included in this model, gender was the only time-independent covariate, as its value did not change over time. Age and the indicator variables for survey round were treated as Type I time-dependent covariates, as their values changed over time but involved no feedback associated with the response; the value for age was known based on the initial value of age and the survey round, and the indicators for survey round were the same for all children. Following the results of the hypothesis test for TDC type by Lai and Small (2007), BMI was treated as a Type II TDC.

As there was no "true" model for empirical data, there was no model to which all others were referenced. Instead, the proposed fit statistics were constructed for the candidate models to try and evaluate the most "ideal" model for these data.

A model with five potential predictors implied that there were $\binom{5}{1}$ unique models that included only one predictor, $\binom{5}{2} = 10$ unique models that included exactly two predictors, $\binom{5}{3} = 10$ unique models that included exactly three predictors, and $\binom{5}{4} = 5$ unique models that included four predictors. Including the full model represented by Equation (16), there was a total of $5 + 10 + 10 + 5 + 1 = 31$ potential models possible for this data situation.

The Process for the Filipino Child Mortality Data

The Kullback-Leibler Information Criterion (KLIC) was calculated for all 31 potential models to predict future morbidity of the Filipino children. The model with the

smallest value of the KLIC was considered the most "ideal" model for these data. Those that had KLIC values close to that of the most ideal model fit the data relatively well, and models associated with larger values of the KLIC indicated poorer fit.

The process underlying the use of the chi-squared statistic involved greater challenge. As this statistic was used to compare two candidate models at a time, 31 potential models would imply the pairwise comparisons of $\binom{31}{2} = 465$ total models. Additionally, a Bonferroni-adjustment for the traditional Type I Error of $\alpha = 0.05$ would require an adjusted significance level of $\alpha^* = 0.05/465 = 1.075 \times 10^{-4}$, which was unreasonable. To avoid the tediousness and issues posed by 465 comparisons, the results of the process involving the Kullback-Leibler Information Criterion was used to address this challenge. After the KLIC values were obtained for the 31 potential models, the models with the 5 smallest values of the KLIC were set aside. The model selected as most ideal using the KLIC was compared to the second, third, fourth, and fifth most ideal models, for a total of four comparisons. This helped reduce the task from 465 pairwise comparisons to 4 comparisons, and the Bonferroni-adjustment only required α to be adjusted by a factor of 4, or $\alpha^* = 0.05/4 = 0.0125$.

Results to be Reported: The Filipino Child Mortality Data

After the estimated values of the Kullback-Leibler Information Criterion and the chi-squared statistics for comparing two candidate models had been obtained for the models mentioned in the previous sections, several critical pieces of information were reported. This section provides further details.

The estimated Kullback-Leibler Information Criterion for all 31 models was reported, and the most "ideal" model was selected for interpretation. Models that had the

estimated KLIC values close to that of the ideal model were briefly examined to weigh parsimony into the selection of the most ideal model. Then, the results of these models were also compared to those obtained in the analysis by Lai and Small (2007).

After the Kullback-Leibler Information Criterion estimates had been obtained, these results were used to select the top five most ideal models, namely, those that had the five smallest values of the estimated KLIC. Using the chi-squared statistic, a total of four pairwise comparisons were made between the model identified as most ideal using the KLIC with the remaining four candidate models. The p-values of the chi-squared tests, which are used to determine whether the null hypothesis of the chi-squared test was rejected, were reported.

The Simulation

The ability of the proposed statistics in assessing model fit could only be evaluated once using real data. The use of a simulation study enabled further assessment of the performance of these statistics. Multiple sets of data could be randomly generated in a simulation and the same models built for each set, yielding slightly different values of the fit statistics for each set. As the conditions underlying the data generation were identical, it was expected that model parameters and fit statistics assumed different values, but the overall results of the evaluation of model fit followed a similar pattern (e.g., the true model should often be selected as the most “ideal” model of all candidate models). The focus of this research study was not on the simulation itself; rather, a simulation should support the results of the performance of the proposed fit statistics.

This simulation would have certain characteristics shared with those of the Filipino Child Mortality Data. The real data contained observations from 370 children at

3 different time periods, for a total of 1,110 observations. As this was a relatively small sample size, two different sample size conditions, small sample and large sample cases, were examined in this simulation study.

Simulation Data

Two different sample size cases were examined in this simulation study to evaluate the quality of performance of the proposed statistics in assessing adequate and poor fit for both small and large sample size situations. Following the simulation studies by Lai and Small (2007) and Hosmer et al. (2013), the small sample was comprised of 100 subjects, and the large sample included 500 subjects. Maintaining a balanced structure with $T = 5$ observations per subject for each of the sample size conditions, the small sample case was comprised of $I = 100$ subjects with $T = 5$ repeated observations each for a total of $N = 500$ observations, and the large sample case was comprised of $I = 500$ subjects with $T = 5$ repeated observations each for a total of $N = 2{,}500$ observations (Hosmer et al., 2013; Lai & Small, 2007). Additionally, 2,000 replicates were simulated for each sample size, following the simulation study by Lai and Small (2007).

The simulation data included five predictor variables to keep the data structure consistent with the Filipino Child Mortality Data. A binary response was simulated, as the focus of this research was on the use of binary logistic regression models. However, due to the nature of the occasional poor performance of binary responses modeled with 2-Step Generalized Method of Moments and issues involving non-convergence of the GMM algorithm seen with binary responses in some data situations (Kleiber & Zeileis, 2008), a continuous response was also simulated, and multiple regression models built. The use of linear regression models averted the occasional non-convergence issue,

ensuring an appropriate evaluation of the quality of the proposed statistics in assessing model fit.

All five predictors were continuous, as binary and categorical predictors may pose potential threats of GMM algorithm non-convergence (Kleiber & Zeileis, 2008; Shane, 2013). Since time-dependent covariates of Types II and III were the most challenging to work with in longitudinal data analyses (Lai & Small, 2007), these types of covariates were included in the randomly generated data. Moreover, to keep the properties of the data structure as consistent as possible with the Filipino Child Mortality Data, TDCs of Types I, II, and III were the three types included in the simulation.

Unlike modeling real data, a "true" model existed in situations involving simulations because the data were simulated to possess certain properties; the "true" model was defined by the conditions present in the simulated data characteristics. For each scenario, five models were examined: a "true" model, an underfit model missing an essential Type II TDC, an underfit model missing an essential Type III TDC, an overfit model that included an additional and unnecessary Type II TDC, and an overfit model that included an additional Type III TDC. The models under investigation focused on the incorrect omission or unnecessary inclusion of Type II and Type III time-dependent covariates because TDCs of Types II and III had been known to involve the most challenge in analyses (Lai & Small, 2007).

Models for the Simulation Data

The systematic component for the true model was defined by:

$$\eta_{it} = \beta_0 + \beta_1 x_{it,1} + \beta_2 x_{it,2} + \beta_3 x_{it,3} \ , \tag{17}$$

where $x_{it,1}$ was a continuous time-dependent covariate of Type I, $x_{it,2}$ was a continuous time-dependent covariate of Type II, $x_{it,3}$ was a continuous time-dependent covariate of Type III, η_{it} was the link function that linked the response with the systematic component of the model, and y_{it} was the response value for subject i at time t. When y_{it} was binary, multiple logistic regression models were constructed using the logit-link function; when y_{it} was continuous, multiple linear regression models were constructed using the identity link function.

Two additional unnecessary predictors were randomly generated, a continuous Type II TDC ($x_{it,4}$) and a continuous Type III TDC ($x_{it,5}$). The inclusion and omission of these predictors, as well as the three predictors in the true model, would be helpful in assessing the performance of the proposed fit statistics in assessing poor and adequate fit of candidate models.

A total of five models were constructed as part of this simulation: a true model, two overfit models, and two underfit models. The true model was defined by Equation (17) above and were referred to as model M_0. The first underfit model, which was denoted M_{U1}, excluded the Type II TDC in Equation (17):

$$M_{U1} : \eta_{it} = \beta_0 + \beta_1 x_{it,1} + \beta_3 x_{it,3} \ ,$$

and this model illustrated how the proposed fit statistics were influenced when a necessary Type II TDC was missing from a model. The second underfit model, M_{U2}, excluded the Type III TDC in Equation (17):

$$M_{U2} : \eta_{it} = \beta_0 + \beta_1 x_{it,1} + \beta_2 x_{it,2} \ ,$$

and this model illustrated how the proposed fit statistics were influenced when a necessary Type III TDC was missing from a model.

The first overfit model, denoted M_{O1}, included an additional unnecessary Type II TDC:

$$M_{O1} : \eta_{it} = \beta_0 + \beta_1 x_{it,1} + \beta_2 x_{it,2} + \beta_3 x_{it,3} + \beta_4 x_{it,4},$$

and this model illustrated how the proposed fit statistics were influenced when an unnecessary Type II TDC was added to the true model. Lastly, the second overfit model, M_{O2}, included an additional unnecessary Type III TDC:

$$M_{O2} : \eta_{it} = \beta_0 + \beta_1 x_{it,1} + \beta_2 x_{it,2} + \beta_3 x_{it,3} + \beta_5 x_{it,5},$$

and this model illustrated how the proposed fit statistics were influenced when an unnecessary Type III TDC was added to the true model.

Additionally, true values for the parameters, $\boldsymbol{\beta}$, had to be specified in order to generate data. To align the structure and content of the simulation data with the analysis of the real-world data on Filipino children, the results of Lai and Small (2007) were used: β_0 = -0.580, β_1 = 0.049, β_2 = -0.010, β_3 = 0.091, β_4 = 0.280, and β_5 = -0.004.

The Process for the Simulation Data

The Kullback-Leibler Information Criterion (KLIC) was calculated for all five models for each set of the simulated data. It was expected that the true model was associated with the smallest value of the KLIC, and models with poor fit were associated with larger values of the KLIC. Assuming the performance of this information criterion to be similar in nature to that of the Akaike Information Criterion (AIC) and corrected AIC, it was anticipated that the KLIC would possess the ability to evaluate the poor fit of

underfit models more effectively than that of overfit models, relative to the true model. Hence, the KLIC values of the underfit models could be expected to be slightly larger than those of the overfit models.

The comparison of the fit of two candidate models was used to evaluate the relative fit of each of the underfit and overfit models to that of the true model: M_0 versus M_{U1}, M_0 versus M_{U2}, M_0 versus M_{O1}, and M_0 versus M_{O2}. Further interest was invested in comparing the fit of models missing a necessary type of time-dependent covariate with that of a model overfitted with the same type of TDC. Thus, additional comparisons were made between the underfit model lacking a necessary Type II TDC with the model overfit with an unnecessary Type II TDC, M_{U1} vs. M_{O1}, as well as the underfit model lacking a necessary Type III TDC with the model overfit with an unnecessary Type III TDC, M_{U2} vs. M_{O2}. Thus, a total of six comparisons were made using the statistic similar to the chi-squared test for assessing candidate model fit for each set of the simulated data.

Results Reported for the Simulation Data

As there were two sample size conditions with 2,000 replicates each (Lai & Small, 2007), as well as both binary and continuous response variables, individual estimates of the Kullback-Leibler Information Criterion were not reported. Instead, collapsed and combined results were reported for all comparisons, separately for the two response variable types.

For each sample size, the averages of the estimated KLIC values for each of five models were reported and represented in the form of a bootstrapped 95% confidence interval plot. Confidence interval limits were also reported. In other words, the 2,000 KLIC estimates for the true model were averaged, and the middle 95% were reported.

The same was also computed for the four underfit and overfit models. These five averages and corresponding limits were then plotted. This report was generated for both the small and large sample cases, as well as for binary and continuous responses.

In a similar manner, information obtained from the comparison similar to the chi-squared test were reported for each of the 2,000 runs of each of the two sample size conditions. Six pairwise comparisons were made, as listed in the section "The Process for the Simulation Data." For each pairwise comparison, the proportion of times that the overfit or underfit model was selected over the true model was reported.

As the true model was known and fully specified in a simulation study, it was also possible to obtain information about detection and non-detection proportions using the pairwise comparisons of the underfit and overfit models with the true model. Comparisons of the underfit models with the true model, M_{U1} vs. M_0 and M_{U2} vs. M_0 – two of the comparisons described in the section "The Process for the Simulation Data"--gave information about the detection proportion, or the proportion of times that an incorrect predictor was correctly detected by the KLIC. Comparisons of the overfit models with the true model, M_{O1} vs. M_0 and M_{O2} vs. M_0--two other comparisons described in the previous section--gave information about non-detection proportion, or the proportion of times in the simulation that an incorrect predictor was not detected by the KLIC. The estimated non-detection proportion, based on all 2,000 runs, were reported for the pairwise comparisons of both M_{O1} with M_0 and M_{O2} with M_0 for the small sample and large sample conditions. Information about the detection proportion were reported as the proportion of times the null hypothesis was rejected from all 2,000 runs of the pairwise comparisons of both M_{U1} with M_0 and M_{U2} with M_0 for the small sample and

large sample conditions. To assess the KLIC's detection proportion, or its ability to correctly detect an inappropriate predictor, the traditional power threshold of 80% was considered acceptable for continuous responses that were correlated over repeated observations; however, a reduced power of 50% was more reasonable and was used instead of the traditional 80% for binary correlated responses (Lin & Myers, 2006; Pan, 2001b).

It was expected that both methods, using the KLIC and the comparison similar to the chi-squared test, would yield similar results, selecting the true model as the most ideal model in almost all of the 2,000 replicates for each response type and sample size condition. Discussions of these findings are provided in Chapter IV.

CHAPTER IV

RESULTS

This research investigated a measure to assess the fit of nested models estimated using the Generalized Method of Moments (GMM) in the presence of time-dependent covariates (TDCs). This research was necessary, as the selection process of candidate nested models has been limited to a pairwise chi-squared test when the estimation procedure is moment-based and TDCs are present.

Three research questions were posed in this research, and the goal of answering them was to understand the properties of the Kullback-Leibler Information Criterion (KLIC) as an information criterion to select among candidate GMM models in the presence of TDCs. The KLIC statistic was derived in Chapter III, and it was represented by Equation 14:

$$KLIC = \min_{\gamma} \left(\frac{1}{N} \sum_{i=1}^{N} \exp\left(\hat{\gamma}^T \, \hat{g}_i \left(\hat{\beta} \right) \right) \right),$$

where N denotes the total number of subjects, g_i is the moment condition of the i^{th} subject estimated using $\hat{\boldsymbol{\beta}}$ from 2-Step GMM, and $\hat{\gamma}$ is an unknown parameter estimated by the minimization of Equation (12) using $\hat{\boldsymbol{\beta}}$ from 2-Step GMM.

The three research questions investigated in this study and their corresponding answers were:

Q1 How can information associated with the fit of model parameters estimated using the Generalized Method of Moments be expressed or measured?

A1 The Kullback-Leibler Information Criterion can be used to select among candidate models when data containing time-dependent covariates were modeled using Generalized Method of Moments. This was an alternative to the current chi-squared test of pairwise model comparison.

Q2 What is the detection proportion of the model selection process of such measures in their ability to detect poor fit of underfit models?

A2 The Kullback-Leibler Information Criterion had moderate to strong ability to detect the poor fit of underfit models. The detection proportion of the KLIC to select the correct model over the underfit model ranged from approximately 60-80% for binary outcomes and approximately 85-90% for continuous outcomes.

Q3 What are the non-detection proportions of such measures in suggesting poor fit for appropriate models?

A3 The KLIC suffered excessive non-detection proportions in its verdict of poor fit for appropriate models against an overfit model. Non-detection proportions for binary data models were as high as 35% when an extraneous Type II TDC was introduced to the correct model and spiked to almost 90% when the extraneous TDC was of Type III. Non-detection proportions for continuous data models were reasonable, below 10%, when the extraneous predictor was a Type II TDC and as high as 40-50% when the extraneous predictor was a Type III TDC.

The answers to these research questions are explained in greater detail in this chapter, including the details of the KLIC's detection proportion to identify the poor fit of underfit models and its occasional indication of poor fit for appropriate models.

The performance of the proposed Kullback-Leibler Information Criterion (KLIC) statistic was evaluated using both the analysis of real data and a simulation. The Filipino Child Mortality (FCM) data set was used as an example of model fitting, and the purpose of testing the KLIC on real data was to assess its performance in the comparison of every possible nested model. A simulation involving randomly-generated data containing TDCs was also used to evaluate the performance of the KLIC, as this provided additional data sets with known conditions and parameters, which could not be stipulated for real data.

Analysis of the Filipino Child Mortality Data

For the analysis of the Filipino Child Mortality (FCM) data, two cases were examined: a case with a binary response and a case with a continuous response. For each case, the analyses conducted are described below as well as in the previous chapter. The binary response case was analyzed first, with findings and discussions provided in the "Binary Response Analysis" section. The continuous response case applied the log-transformation provided in Equation 15 to the binary outcome. Results and discussions of the analysis of the continuous outcome data are included in the "Continuous Response Analysis" section.

Binary Response Analysis

For the analysis of the binary data, 31 unique models were estimated and compared with respect to model fit. These 31 models were described in the sections titled "The Models for the Filipino Child Mortality Data" and "The Process for the Filipino Child Mortality Data" in Chapter III. The models to predict morbidity was defined by Equation 16 in Chapter III:

$$\eta_{it} = \beta_0 + \beta_1 x_{it,\,Age} + \beta_2 x_{it,Gender} + \beta_3 x_{it,\,BMI} + \beta_4 x_{it,\,Round2} + \beta_5 x_{it,\,Round3} + \varepsilon_{it}\,,$$

where $x_{it,Age}$ was the covariate for age of subject i at time t, $x_{it,Gender}$ was the covariate for gender of subject i at time t, $x_{it,BMI}$ was the covariate for BMI of subject i at time t, $x_{it,Round2}$ and $x_{it,Round3}$ were the indicator variables for survey rounds 2 and 3, respectively, of subject i at time t, and ε_{it} was the error term. The systematic component $\eta_{it} = \mu_{it} = p_{it}$ was defined as the probability of success for subject i at time t, and g was the logit-link function.

The Kullback-Leibler Information Criterion (KLIC) was calculated for each model, and the models were ranked based on the KLIC. The model with the smallest KLIC was considered the most ideal candidate model, and the model with the largest KLIC was considered the least ideal candidate model.

The top five models selected by the KLIC, including the value of the KLIC, are listed in Table 1. The KLIC statistics for all 31 models are provided in Table 7 in Appendix B. The most ideal model was the full model. This was not surprising, as the full model has been commonly selected as the most ideal model by most model selection criteria, unless the data contain unrealistic or meaningless predictors. The KLIC method is similar to the AIC method in that it calculates the amount of information lost from fitting an approximation to the observed data and no hypothesis tests are involved. A key difference is that the AIC includes a penalty term for less parsimonious models, and this penalty could sometimes demote the full model from being selected as most ideal. The KLIC method currently does not include a penalty term, so its selection of the full model as most ideal was realistic; it selected the model that lost the least amount of information compared to the original data. In terms of magnitude, the KLIC of the full model was slightly smaller than that of M_2 and was noticeably smaller than the KLIC statistics for models M_3-M_5.

Table 1

Kullback-Leibler Information Criterion of the Top Five Candidate Models for the Binary Data

	Kullback-Leibler Information Criterion (KLIC)	Variables included
Most ideal model (M1)	148.7	Full model
2nd candidate model (M2)	167.9	Model without BMI
3rd candidate model (M3)	202.3	Model without survey round 2
4th candidate model (M4)	251.1	Model without BMI and survey round 2
5th candidate model (M5)	311.4	Model without gender

The second ideal model selected by the KLIC was the model with all of the predictors except for body mass index (BMI). This was not an intuitive finding; rather it was contrary to the notion that morbidity of ill or malnourished children would likely be impacted by his/her BMI. However, this result aligned with the conclusions that Lai and Small (2007) reached in their research--namely, that BMI was not a strong predictor of future morbidity. One possible reason included in their discussion, as well as the discussions of Bouis and Haddad (1990) and Bhargava (1994), was that local conditions may have influenced how BMI affected morbidity among the sampled children. Another possible explanation was the lack of additional data; the current data were limited in the availability of potential predictors. Morbidity could be affected by a variety of factors, and those external factors were not captured in these data. Examples of these external factors are: nutrition; the average number of meals the child consumes in a given week;

currency on vaccinations (e.g., the measles, mumps, rubella (MMR) vaccine); the number of children or siblings in the household; the number of caregivers in the household, or a ratio of caregivers to children in the household; distance from the closest clinic or hospital; distance from the closest city of population 50,000 or greater; and average monthly income, or the ratio of average monthly income to the number of adults and children in the family (Sommer, Katz, & Tarwotjo, 1984; Sommer, Tarwotjo, Hussaini & Susanto, 1983; Sommer, Tarwotjo, & Katz, 1987).

The third candidate model selected by the KLIC was the model without the indicator for survey round 2. A possible explanation for the omission of the indicator for survey round 2 was that the data collected at survey round 3 were more impactful on child morbidity than those collected at the prior iteration--more impactful to the extent that survey round 3, without 2, sufficiently predicted childhood mortality. The health conditions of children who did not survive could have deteriorated more significantly between the first and third survey visitations, rather than between the first and second survey visitations. On the contrary, the health conditions of children who survived may have improved by the third and final survey round. For children whose health conditions remained relatively stable throughout the 4-month study, there would not have been a noticeable spike or decline throughout the three survey rounds. In a sense, the data collection dates were too close in proximity to make survey round 2 a meaningful predictor of morbidity.

The fourth candidate model selected by the KLIC method was the model without BMI and the indicator for survey round 2. The realizations made previously regarding the weak impact of BMI on child morbidity justified the selection of another model without

BMI. Similarly, the same discussions and justifications provided for the selection of model M_3 could be applied to the selection of M_4 as the fourth candidate model. The only difference was that M_3 included BMI as a predictor. The KLIC method, like AIC, calculated the amount of information lost by approximating the full data using a model. The inclusion of BMI in candidate model M_3 could have been due to slightly more information becoming gained by including BMI in the model. Although BMI did not appear to be a meaningful predictor of child morbidity in this analysis, the amount of information lost by excluding both BMI and the indicator for survey round 2 was more than the amount of information lost by excluding only the indicator for survey round 2.

The fifth candidate model selected by the KLIC was the model without gender. From a conceptual standpoint, the gender of children should not have a substantial effect on morbidity. The survivability of children who were very ill, sufficiently ill to the extent of mortality, should not be affected by the child's gender. In this sense, it made sense that the model without gender was selected as one of the top five candidate models in predicting mortality.

All of the top five models selected by the KLIC statistic included age and survey round 3. Age was an important predictor of morbidity. It was not a surprise that for children between 1 and 14 years of age, the age at which he/she would get very ill could considerably affect mortality. For children in this age range, it made sense that age impacted mortality more than the child's gender or survey round.

Survey round 3 was also included in the top five models selected by the KLIC. The same argument provided previously, in the discussion of the selection of M_3 as the third candidate model, applied. It was possible that the data collected during the final

visit, survey round 3, was more indicative of the child's overall health condition, and hence mortality, than those collected at the earlier surveys. The surveys were conducted only 4 months apart, so it was possible that there was not sufficient change in the overall health of children between survey rounds 1 and 2, while there was discernable change between survey rounds 1 and 3.

The values of the KLIC statistics obtained for the 31 models varied greatly, from a value in the low hundreds to the high hundreds. The range of KLIC values was roughly 400. As models got worse, the KLIC grew larger, and there were a few big jumps. The most notable increases in the KLIC values happened at the tail ends of the models: among the top five candidate models and among the bottom five candidate models. The KLIC increased by about 20 between the full model, which was selected as the most ideal model, and the second-best model. The increase was even more rapid between the next few candidate models. It increased by approximately 35 between the second and third models and by 50 and 60 between the third and fourth models and fourth and fifth models, respectively. Similarly, for the five least ideal models, the KLIC increased by approximately 20 between 2 consecutive models. The only exception in this trend was for the models that contained only the indicator for survey round 2 and only the indicator for survey round--the KLICs of these models were very similar. This finding implied that the indicators for survey rounds 2 and 3 explained approximately the same amount of information in the response.

Chi-squared tests for the binary response analysis. The chi-squared method of model comparison was used to take a closer look at the models estimated for the FCM

data. The goodness-of-fit of two nested GMM models, M_1 and M_2, could be compared using the statistic in Equation 11:

$$C\left(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2\right) = \left(\min\left\{ QF\left(\hat{\boldsymbol{\beta}}_1\right)\right\} - \min\left\{ QF\left(\hat{\boldsymbol{\beta}}_2\right)\right\}\right) \sim \chi^2_{(p_1 - p_2)},$$

where QF($\hat{\boldsymbol{\beta}}_1$) and QF($\hat{\boldsymbol{\beta}}_2$) are the quadratic forms of the two nested models, $\hat{\boldsymbol{\beta}}_1$ is the vector of parameters estimates obtained under M_1, $\hat{\beta}_2$ is the vector of parameters estimates obtained under M_2, and p_1 - p_2 denotes the difference in the number of parameters between the two candidate models.

The difference between the minimum of the GMM quadratic forms of two competing models follows a chi-squared distribution with degrees of freedom equal to the difference between the numbers of parameters in the two models (Hansen et al., 1996), and this statistic can be used to test the hypothesis:

H_0: The candidate model with fewer parameters is sufficient in explaining the information in the response.

H_A: The candidate model with fewer parameters is *not* sufficient in explaining the information in the response.

It was not feasible to compare all possible pairs of the 31 models, so 4 chi-squared comparisons were considered: the most ideal model selected by the KLIC, the full model (M_1), was compared to the second, third, fourth, and fifth candidate models selected by the KLIC, models M_2-M_5, respectively. The chi-squared comparisons of these models tested the null hypothesis that the candidate model with fewer parameters, model M_j, was sufficient in explaining the information in the response, for j = 2, 3, 4, 5. More specifically, the null hypotheses tested in these comparisons were:

H01: Model M_2 explains sufficient information in the response

H02: Model M_3 explains sufficient information in the response

H03: Model M_4 explains sufficient information in the response

H04: Model M_5 explains sufficient information in the response

The *p*-value of the chi-squared tests was compared to $\alpha^* = 0.0125$ instead of the notional $\alpha = 0.05$, as four comparisons were made. A *p*-value smaller than 0.0125 suggested that model M_1 explained significantly more information in the data relative to the compared model, M_j.

The chi-squared comparisons of the top five models selected by the KLIC, including the *p*-value of the chi-squared test, are provided in Table 2. The first column of the table denotes the model that was compared to the most ideal model. Overall, the results of the chi-squared tests concurred with the results of the KLIC method, which selected the full model as the most ideal model for these data. The full model outperformed the more parsimonious second, third, fourth, and fifth candidate models, based on both the KLIC statistic and the chi-squared test of model comparison.

Table 2

The p-values of the Chi-squared Comparisons for the Binary Data Models

Model being compared	p-value of χ^2 comparison	Interpretation
2nd candidate model	0.0108	The full model fit the data better than the 2nd candidate model
3rd candidate model	0.0017	The full model fit the data better than the 3rd candidate model
4th candidate model	0.0004	The full model fit the data better than the 4th candidate model
5th candidate model	< 0.0001	The full model fit the data better than the 5th candidate model

Although the *p*-values for all four comparisons were smaller than 0.0125, the *p*-value of the comparison between the full model (M_1) and model M_2 was fairly close to 0.0125. This suggested that model M_2 fit the data almost as well as the full model, and its fit was only slightly inferior to that of the full model. However, multiple chi-squared comparisons were made, so the *p*-values of these tests were compared to 0.0125 instead of the notional significance level of 0.05. If only one comparison had been made, namely, the comparison between the fits of models M_1 and M_2, and the *p*-value of this test had been compared to the standard threshold of 0.05, model M_1 fit the data substantially better than model M_2. Therefore, from a practical standpoint, model M_1 clearly outperformed model M_2. This was verified by comparing the values of the KLIC between models M_1 and M_2; the difference was noticeable. Furthermore, the most ideal model, model M_1, included all predictors, while model M_2 did not include BMI as a predictor of morbidity. The inclusion of BMI in the most ideal model suggested that BMI contained enough information in accurately fitting these models to the data. This should not be confused with the discussion in the previous section; the frequent omission of BMI in the top five candidate models meant that BMI was a weak predictor of morbidity for these data. However, model fit peaked when all predictors were included. These were not contradictory statements but refer to different aspects of model estimation. The inclusion of all predictors lost the least amount of information available in the full data. The amount of information lost by omitting BMI in the GMM estimation was equivalent to the difference in the KLIC values between models M_1 and M_2.

The *p*-value associated with the comparison between models M_1 and M_3 suggested that the most ideal model selected by the KLIC, the full model, outperformed

model M_3. A comparison of the KLIC values between models M_1 and M_3 also supported this finding, as the difference was not trivial. While the most ideal model included all predictors, model M_3 was missing the indicator variable for survey round 2, which suggested that survey round 2 was a key piece of information in estimating child mortality rates. The amount of information lost by omitting the indicator for survey round 2 in the GMM estimation was equivalent to the difference in the KLIC values between models M_1 and M_3.

Based on the magnitude of the *p*-values associated with the comparison between models M_1 and M_4, as well as the comparison between models M_1 and M_5, the full model outperformed both models M_4 and M_5 significantly. These comparisons indicated the importance of the variables included in the full model, M_1, and the effects on model fit when a meaningful variable was omitted. The inflation in the KLIC value of models M_4 and M_5, compared to the KLIC of model M_1, conveyed the importance of the variables included in M_1 that were missing in models M_4 and M_5.

Continuous Response Analysis

For the transformed, continuous response data, all 31 models were also estimated and compared with respect to model fit. For the analysis of the continuous data, 31 unique models were estimated and compared with respect to model fit. These 31 models were described in the sections titled “The Models for the Filipino Child Mortality Data” and “The Process for the Filipino Child Mortality Data” in Chapter III. The models to predict morbidity was defined by Equation 16 in Chapter III:

$$\eta_{it} = \beta_0 + \beta_1 x_{it,\,Age} + \beta_2 x_{it,\,Gender} + \beta_3 x_{it,\,BMI} + \beta_4 x_{it,\,Round2} + \beta_5 x_{it,\,Round3} + \varepsilon_{it},$$

where $x_{it,Age}$ was the covariate for age of subject i at time t, $x_{it,Gender}$ was the covariate for gender of subject i at time t, $x_{it,BMI}$ was the covariate for BMI of subject i at time t, $x_{it,Round2}$ and $x_{it,Round3}$ were the indicator variables for survey rounds 2 and 3, respectively, of subject i at time t, and ε_{it} was the error term. The systematic component $\eta_{it} = \mu_{it}$ was defined as the mean response for subject i at time t, and g was the identity link function.

The Kullback-Leibler Information Criterion (KLIC) was calculated for each model, and the models were ranked based on the KLIC. The model with the smallest KLIC was considered the most ideal model, and the model with the largest KLIC was considered the least ideal of the remaining 30 models.

The top five models selected by the KLIC, including the value of the KLIC are listed in Table 3. The KLIC statistics for all 31 models are provided in Table 8 in Appendix C. The most ideal model was, again, the full model. This was not a surprise, as the full model was selected by the KLIC statistic as the most ideal model for the binary data analysis. Mathematically, the full model is oftentimes selected as the most ideal model because it utilizes all of the information in the data--namely, all of the variables. Per the previous discussion, using the KLIC as a model selection criterion was a novel method, and it currently does not include a penalty term, similar to those of the AIC or BIC. With respect to practical model selection, the KLIC of the most ideal model and the second most ideal model were very similar. They were essentially equally good at predicting mortality.

Table 3

Kullback-Leibler Information Criterion of the Top Five Candidate Models for the Continuous Data

	Kullback-Leibler Information Criterion (KLIC)	Variables included
Most ideal model (M_1)	287.9	Full model
2nd candidate model (M_2)	293.5	Model without BMI
3rd candidate model (M_3)	320.3	Model without survey round 2
4th candidate model (M_4)	324.1	Model without survey round 3
5th candidate model (M_5)	376.6	Model without gender

The KLIC selected the model without BMI as the second most ideal model, similar to the results of the binary outcome analysis. Although this was not intuitive at first, the same arguments from earlier were applied here--for this sample of children in the Philippines, BMI did not predict their mortality or survivability well. There could have been external factors that influenced the effect of BMI on mortality, such as the family's socioeconomic class and nutrition. If parsimony was an important factor in the estimation process, the second candidate model was almost equally as good at predicting mortality, as the omission of BMI did not change the KLIC by much.

The third candidate model selected by the KLIC was the model without the indicator for survey round 2, similar to the findings of the binary data analysis. The same explanation applied here: the data collected at survey round 3 were more impactful on child morbidity than those collected at the prior iteration--more impactful to the extent

that survey round 3, without survey round 2, sufficiently predicted childhood mortality. It could be possible that the change in health conditions of children was more noticeable between survey rounds 1 and 3, or there was not enough time between survey rounds 1 and 2, or between survey rounds 2 and 3, for a detectable change.

The fourth candidate model selected by the KLIC method was different for the continuous and binary data analyses. For the binary outcome data, the KLIC selected the model without BMI and survey round 2. For the transformed response data, the KLIC selected the model without the indicator for survey round 3. One possible explanation was that the indicator variables for survey rounds 2 and 3 affected child mortality almost equally, with the difference that the model with survey round 3 did so slightly better than the model with survey round 2. Holding all other variables constant, the inclusion of the indicator for survey round 3 provided a slightly better prediction of child mortality than including the indicator for survey round 2. However, the difference was negligible, as their KLIC values were very similar.

The fifth candidate model selected by the KLIC was the model without gender. This lined up with the results of the binary data analysis, in which the model without gender was also selected as the fifth candidate model. Using the same rationale from the section on “Binary Response Analysis,” the gender of children should not have a substantial effect on morbidity.

Overall, the KLIC selected the model that included all five variables (i.e., the full model) as the most ideal model, and the second to fifth candidate models included four of the five variables from the data set. These results suggested that the KLIC was more sensitive to the number of variables included in the model for continuous outcome data.

Further research to consider the addition of a penalty term to the KLIC may shed more light on the performance of the KLIC as a model selection criterion for the analysis of continuous outcome data involving time-dependent covariates. This is discussed in Chapter V.

Similar to the binary outcome analysis, age was included in all of the top five models selected by the KLIC, which suggested that it was an important predictor of morbidity for this sample of children. Again, this was no surprise; for children between the ages of 1 and 14 years in this geographic area, it made sense that the child's age impacted mortality more than his/her gender or survey round.

The values of the KLIC statistics obtained for the 31 models varied greatly, from a value in the low hundreds to the high hundreds. The range of KLIC values was roughly 500. As models got worse, the KLIC grew larger, and there were a few big jumps. The most notable increases in the KLIC values happened when a meaningful predictor, such as age, was lost in a subsequent model. The KLIC increased by at least 15-20 when consecutive models were alike, except for the omission of age as a predictor. This suggested that age was an important predictor that explained a substantial amount of information in child morbidity. On the contrary, the change in KLIC was trivial for two similar models in which the only difference was whether the dummy variable was for survey round 2 or 3. This suggested that the indicator variables for survey rounds 2 and 3 were essentially exchangeable, in the sense that they explained the same amount of information in child morbidity.

Chi-squared tests for the continuous response analysis. The chi-squared method of model comparison was used to take a closer look at the continuous response

models estimated for the FCM data. Again, it was not feasible to compare all possible pairs of the 31 models, so the 4 most meaningful chi-squared comparisons were conducted. The most ideal model selected by the KLIC, the full model (M_1), was compared to the second, third, fourth, and fifth candidate models selected by the KLIC, models M_2-M_5, respectively.

The chi-squared comparisons of these models tested the null hypothesis that the candidate model with fewer parameters, model M_j, was sufficient in explaining the information in the response, for j = 2, 3, 4, 5. More specifically, the null hypotheses tested in these comparisons were:

H01 Model M_2 explains sufficient information in the response

H02 Model M_3 explains sufficient information in the response

H03 Model M_4 explains sufficient information in the response

H04 Model M_5 explains sufficient information in the response

The *p*-value of the chi-squared tests was compared to $\alpha^* = 0.0125$ instead of the notional $\alpha = 0.05$, as four comparisons were made. A *p*-value smaller than 0.0125 suggested that model M_1 had a significantly better fit relative to the compared model, M_j.

The chi-squared comparisons of the top five models selected by the KLIC, including the *p*-value of the chi-squared test, are provided in Table 4. The first column of the table denotes the model that was compared to the most ideal model. Overall, the most ideal model selected by the KLIC outperformed the second, third, fourth, and fifth candidate models. The results of the chi-squared tests concurred with the results of the KLIC method, which selected model M_1, the full model, as the most ideal model for these

data. A big portion of these results also aligned with the analysis of the binary outcome data.

Table 4

The p-values of the Chi-squared Comparisons for the Continuous Data Models

Model being compared	*p*-value of χ^2 comparison	Interpretation
2nd candidate model	0.0119	The full model fit the data better than the 2nd candidate model
3rd candidate model	0.0087	The full model fit the data better than the 3rd candidate model
4th candidate model	0.0044	The full model fit the data better than the 4th candidate model
5th candidate model	0.0009	The full model fit the data better than the 5th candidate model

The *p*-value of the comparison between the full model (M_1) and model M_2 was fairly close to the Bonferroni-adjusted significance level of 0.0125. This suggested that model M_2 fit the data almost as well as the full model, and its fit was only slightly inferior to that of the full model. However, had there not been four comparisons, the *p*-value of this comparison would have been compared to the standard significance level of 0.05, it would have been possible to claim that model M_1 fit the data substantially better than model M_2. In this sense, the full model outperformed model M_2 significantly. This was verified by comparing the values of the KLIC between models M_1 and M_2; the difference was noticeable.

The *p*-values of the comparisons of candidate models M_3 and M_4 to the full model were different by only 0.0043, which was almost negligible. Based on this finding, the

model fit of M_3 relative to the full model was approximately the same as the model fit of M_4 relative to the full model. In other words, the amount of information lost by fitting model M_3 instead of the full model was approximately the same as the information lost by fitting model M_4 instead of the full model. Again, this was verified by comparing the values of the KLIC statistics between models M_3 and M_4; the difference was trivial.

Based on the magnitude of the *p*-values associated with the comparison between models M_1 and M_5, the full model outperformed model M_5 significantly. This comparison indicated the importance of the variables included in the full model, M_1, and the effects on model fit when a meaningful variable was omitted. The inflation in the KLIC value of model M_5, compared to the KLIC of model M_1, conveyed the amount of information lost by omitting gender as a predictor of morbidity.

The *p*-values of the chi-squared tests for pairwise comparisons of the top five models were smaller for the binary outcome analysis than they were for the continuous outcome analysis. Also, the *p*-values for the continuous outcome analysis were closer in value to each other than those for the binary outcome analysis. These two observations indicated that the top five models selected by the KLIC in the continuous outcome analysis were deemed to be more similar with respect to the amount of information explained. One possible explanation was that the continuous outcome captured more information (for all predictors) than the binary outcome. The dichotomization of binary outcome data lost enough information that the differences among the top five models in the binary outcome analysis were more drastic than was the case for the continuous outcome analysis.

Simulation

Data Generation Process

Data were simulated according to the algorithm described in this section. The longitudinal response vector $\boldsymbol{Y}$ was generated according to a distribution D, with mean $\boldsymbol{\mu_Y}$ modeled using known predictors $\boldsymbol{\beta}$, summarized as the generalized linear model (McCullagh & Nelder, 1989):

$$\boldsymbol{Y} \sim D\left(\boldsymbol{\mu_Y}, \boldsymbol{Var}(\boldsymbol{\mu_Y})\right),$$

$$g(\boldsymbol{\mu_Y}) = \boldsymbol{X\beta}\ ,$$

where $\boldsymbol{Y}$ indicates the response vector with mean $\boldsymbol{\mu_Y}$ and variance-covariance structure $\mathbf{Var}(\boldsymbol{\mu_Y})$, g indicates the link function, $\boldsymbol{X}$ the design matrix, and the parameter vector $\boldsymbol{\beta}$ included both time-independent covariates and time-dependent covariates (TDCs). For the binary response simulation, D was the binomial distribution, $\mu_{Y_{it}} = p_{it}$ was the probability of success for subject i at time t, $Var\left(\mu_{Y_{it}}\right)$ was calculated as $np_{it}(1 - p_{it})$, and g was the logit link function. For the continuous response simulation, D was Gaussian, $\mu_{Y_{it}}$ was the mean response for subject i at time t, $Var\left(\mu_{Y_{it}}\right)$ was calculated as σ^2, and g the identity link function.

Data were simulated sequentially using auto-regressive terms to represent both autocorrelation and feedback between responses and TDCs. The response mean was simulated according to the following equation:

$$g((\mu_Y)_{it}) = (X\beta)_{it} + \sum_{k=1}^{K} \rho_{XY_k} logit\left(F_{X_k}\left(x_{ik(t-1)}\right)\right) + \sum_{s=t-L}^{t-1} \rho_{YY_s}\, logit\left(F_Y(y_{is})\right), \qquad (18)$$

where K is the number of TDCs, ρ_{XY_k} is the weight of the effect of each TDC $\boldsymbol{x_k}$ on future responses, F_{X_k} is the cumulative distribution function associated with the data type of $\boldsymbol{x_k}$, L is the time-length of the auto-correlation in the responses, ρ_{YY_s} is the weighted effect of each prior response y_{is} on the current response, and F_Y is the cumulative distribution function associated with the data type of the response. The two weighted terms represent contributions from the previous response and are described as follows.

The first weighted term $\rho_{XY_k} logit\left(F_{X_k}(x_{ik(t-1)})\right)$ was the contribution from the previous value of TDC k, and was included for TDCs of Types II and III. For Type I TDCs (and Type IV TDCs), no association existed between prior covariate values and current response values, and so ρ_{XY_k} was set to 0 in those cases. For the sake of future simulations utilizing noncontinuous covariates, the x_{ik} values were transformed to calculate the weight on the scale of the systematic component of the generalized linear model, which was a continuous scale. The logit-CDF transformation accomplished this, as the distribution function F_{X_k} produced a value on the interval (0, 1), whose logit could be any real number. Additionally, the logit-CDF transformation incorporated information from both the mean of $x_{ik(t-1)}$, as well as the actual value generated. In future simulations that may involve binary covariate data, the logit-CDF transformation would preserve information in case binary covariates collapsed to either 0 or 1. Without such a transformation, the effects from prior covariate values could be inconsistent for non-continuous data types. These contributions $\rho_{XY_k} logit\left(F_{X_k}(x_{ik(t-1)})\right)$ were then summed over all TDCs.

The second weighted term $\rho_{YY_s} logit\left(F_Y(y_{is})\right)$ was based on similar logic with respect to the auto-correlation in the response. The weight ρ_{YY_s} represented the strength of the association in responses at different times and could be set to 0 for certain values of *s* to prevent auto-correlation or to restrict the auto-correlation to a known structure -- for example, truncated *AR(1)*. The logit-CDF transformation mapped any type of response data to a continuous value, appropriately incorporated into the systematic component of the data generation model.

Once all components of the right-hand side of Equation 18 were calculated, the mean response for subject *i* at time *t* was calculated by applying the inverse of the link function g^{-1}. Each response value y_{it} was simulated using R version 3.1.0.

Next, TDCs were simulated according to the following equation:

$$h\left(\left(\mu_{x_k}\right)_{it}\right) = \beta_{x_k} \rho_{YX_k} logit\left(F_Y\left(y_{i(t-1)}\right)\right), \quad (19)$$

where *h* is the standard link function for the data type of TDC $\boldsymbol{x_k}$, $\left(\mu_{x_k}\right)_{it}$ is the mean value used for simulation of TDC $\boldsymbol{x_k}$ for subject *i* at time *t*, β_{x_k} is the constant mean for $\boldsymbol{x_k}$, and ρ_{YX_k} is the weight of the effect of the previous response on $\boldsymbol{x_k}$.

The weighted term $\rho_{YX_k} logit\left(F_Y\left(y_{i(t-1)}\right)\right)$ was a contribution from the previous response value and was included for Type III (and Type IV) TDCs. For Type I and Type II TDCs, there was no relationship between prior response values and current covariate values, and so ρ_{YX_k} was set to 0 in those cases. The logit-CDF transformation of the previous response was used to allow for a continuous weight from previous responses of any data type.

Input values for the data generated in this simulation were specified as follows. The vector of true parameters, $\boldsymbol{\beta}$, were set to be equal to the results from Lai and Small (2007): $\beta_0 = 0.580$, $\beta_1 = -0.049$, $\beta_2 = -0.010$, $\beta_3 = -0.091$, $\beta_4 = -0.280$, and $\beta_5 = 0.004$. Values of σ associated with these parameters for the continuous data were: $\sigma_0 = 1$, $\sigma_1 = 2.2$, $\sigma_2 = 3.5$, $\sigma_3 = 1.5$, $\sigma_4 = 4.2$, *and* $\sigma_5 = 0.8$. For the generation of time dependent covariates (TDCs), ρ_{XY} was set to 0.25 for TDCs of Types II and III and 0 for TDCs of Type I, ρ_{YX} was set to 0.25 for TDCs of Type III and 0 for TDCs of Types I and II, and ρ_{YY} was set to 0.25.

Once all components of the right-hand side of Equation 19 were calculated, the mean value of TDC *k* for subject *i* at time *t* was calculated by applying the inverse of the link function h^{-1}. Each covariate value x_{kit} was simulated using R version 3.1.0.

Simulation Cases

Four cases were considered in the simulation analysis:

1. Case 1: Small-sample ($N = 500$) binary-outcome data
2. Case 2: Large-sample ($N = 2{,}500$) binary-outcome data
3. Case 3: Small-sample ($N = 500$) continuous-outcome data, and
4. Case 4: Large-sample ($N = 2{,}500$) continuous-outcome data.

For each case, 2,000 sets of data were generated, with each case including 1 response variable (binary or Gaussian, depending on the case) and 5 explanatory variables: one Type I TDC, two Type II TDCs, and two Type III TDCs. Five models were estimated for each of the 2,000 repetitions of the 4 cases, resulting in the estimation of a total of $5 \times 2{,}000 \times 4 = 40{,}000$ models in the simulation. For each of the 2,000 reps, the 5 models were defined as follows. Further, for the remainder of this research and the discussion

involving the simulation, the models described and compared will be abbreviated and referred to as:

M_0: The "correct" model, which included a Type I TDC, a Type II TDC, and a Type III TDC

M_{U1}: The model that is missing an essential Type II TDC (underfit model 1)

M_{U2}: The model that is missing an essential Type III TDC (underfit model 2)

M_{O1}: The model that has an unnecessary Type II TDC (overfit model 1)

M_{O2}: The model that has an unnecessary Type III TDC (overfit model 2)

Model Kullback-Leibler Information Criterion (KLIC) Averages

The KLIC was obtained for each model, resulting in 2,000 KLICs per model per case, and the average of the KLICs was calculated for each model of a particular case. As 5 models were estimated for each of the 2,000 iterations within a case, this resulted in 5 averaged KLICs per case, for a total of $5 \times 4 = 20$ average KLIC values, which are reported in Table 5:

The average KLIC of model M_0, which was specified to be the "true model," was the smallest of all five models for both of the continuous cases but not for the binary cases. For the binary data simulation, the average KLIC of model M_{O2}, which was the overfit model with an additional Type III TDC, was the smallest of all five models. However, the average KLIC of models M_0 and M_{O2} were very close for the binary cases, which meant that both models fit the data almost equally as well. The average KLIC of model M_{O1}, the overfit model with an additional Type II TDC, was the third smallest of all five models across the board, for both the binary and continuous cases. Similarly, the

two underfit models had the largest average KLIC for both binary and continuous cases, with model M_{U2} outperforming model M_{U1} in each case.

Table 5

Average Kullback-Leibler Information Criterion (KLIC) of the Models Estimated in the Simulation

	Avg KLIC of M_0	Avg KLIC of M_{U1}	Avg KLIC of M_{U2}	Avg KLI of M_{O1}	Avg KLIC of M_{O2}
Case 1: Binary data with small sample size	210.1	246.8 (239.0, 254.6)	231.3 (221.2, 241.4)	214.5 (205.2, 223.8)	207.4 (194.5, 220.3)
Case 2: Binary data with large sample size	201.4	238.7 (232.9, 244.5)	220.6 (213.4, 227.8)	207.9 (200.5, 215.3)	198.2 (185.6, 210.8)
Case 3: Continuous data with small sample size	331.9	396.3 (385.7, 406.9)	379.8 (368.7, 390.9)	356.9 (344.9, 368.9)	334.0 (319.7, 348.3)
Case 4: Continuous data with large sample size	326.3	380.4 (370.2, 390.6)	368.2 (356.3, 380.1)	343.1 (331.4, 354.8)	327.5 (312.4, 342.6)

Note. KLIC is the Kullback-Leibler Information Criterion.
M_0 is the correct model. It has five predictors: 1 Type I TDC, 1 Type II TDC, and 1 Type III TDC.
M_{U1} is the underfit model that is missing an essential Type II TDC.
M_{U2} is the underfit model that is missing an essential Type III TDC.
M_{O1} is the overfit model with an additional Type II TDC.
M_{O2} is the overfit model with an additional Type III TDC.
Values in parentheses are the 95% bootstrapped confidence limits.

Plots of the bootstrapped KLIC statistics of the five models for all four cases are provided in Figures 1, 2, 3, and 4 in Appendix D. These plots served as a visual representation of the variability in the KLIC values for each model in each case. Overall, the distribution of the 2,000 KLIC statistics calculated from each model appeared to be normally distributed across all 4 cases. This made sense, as the KLIC values should be

approximately the same across the 2,000 repetitions of each model, with some random variability around them. This was the case across all 4 cases.

There are several possible explanations for the results of the simulation. The most notable being the performance of the KLIC in selecting the true model as the most ideal model in both the continuous data simulations. This meant that the KLIC was a reasonable model selection criterion for identifying the most ideal model when the response variable was continuous. The differences between the average KLICs of models M_0 and M_{O2} were trivial for both the small sample and large sample cases. A greater distinction may have been observed for much larger sample sizes, such as a sample size of 10,000 or 50,000 instead of 2,000; however, this trivial difference in KLICs was more likely due to the lack of a penalty term for the number of model parameters. The addition of a penalty term may be able to tease out this difference better, allowing the KLIC to more accurately select the true model as the most ideal candidate model, especially for the binary data simulation.

For the binary cases, the KLIC failed to select the true model, M_0, as the most ideal candidate model; however, the average KLIC of model M_0 was not far off from that of M_{O2}, which was selected as the most ideal model. Again, the addition of a penalty term to the KLIC may have allowed it to better identify the true model as the most ideal model for these data. Also, the average KLICs of the two overfit models for the binary cases were not very different, which suggested that the KLIC was not sensitive to distinguishing between overfit models with different types of TDCs. In both cases, it selected the overfit model with an unnecessary Type III TDC as being a better candidate than the overfit model with the unnecessary Type II TDC; however, the difference in

these models' KLICs was trivial. From a practical standpoint, the selection of either model would provide an equally acceptable prediction of the response.

In all four cases of the simulation--both binary and continuous data, as well as small and large sample sizes--the two underfit models, models M_{U1} and M_{U2}, were selected as the least ideal models, and the trend of M_{U1} being selected as the least ideal of the five models was consistent across the board. The consistency of these selections was satisfactory from a simulation perspective, but also slightly challenging to justify initially. The first (and more straightforward) reason was that, again, the KLIC may require a penalty for less parsimonious models. The addition of a penalty term to the KLIC may have helped its ability to select the more parsimonious underfit models over one or both of the overfit models. Also, these results were indicative of the KLIC's sensitivity to the omission of a necessary time-dependent predictor. Model M_{U2} was missing a necessary Type III TDC, and TDCs of this type impart feedback only on the current response value and not on previous or future response values. On the other hand, model M_{U1} was missing a necessary Type II TDC, and TDCs of this type inflict feedback on response values across time. Although both underfit models were missing a necessary time-dependent predictor, the omission of a necessary feedback process may have driven up the KLIC value for model M_{U1}, thus, making it the least candidate model in all four cases of the simulation.

Overall, the KLIC performed slightly better as a model selection criterion for continuous data than binary data, with the caveat that its selection of the true model as the most ideal model was essentially indistinguishable from its selection of one of the overfit models. Similarly, although the KLIC was unable to select the true model as the most

ideal model for the binary data, its selection of model M_{O2} as the most ideal model was also difficult to differentiate from its selection of the true model as the second best model. In this sense, the KLIC's performance achieved practical success with respect to the selection of a leading model among multiple candidate models.

Detection Proportion and Non-Detection Proportion

To further assess the performance of the KLIC, additional comparisons were made. The true model was compared to each of the underfit models, M_{U1} and M_{U2}, to assess the detection proportion of the KLIC's ability to identify the true model as the most ideal candidate model. Similarly, the true model was compared to each of the overfit models, M_{O1} and M_{O2}, to assess the non-detection proportion of the KLIC's performance. For each of these four comparisons, the null hypothesis was: the fit of the true model, M_0, is superior to that of the compared model. The results of these comparisons are tabulated in Table 6:

Discussion of detection proportion. Overall, the detection proportion of the KLIC's ability to select the true model was very high for the comparisons of the true model and underfit models, regardless of the sample size, TDC type, and data type (binary or continuous). In particular, the KLIC showed great performance in selecting the correct model when the outcome was continuous. For the comparisons of M_0 versus both M_{U1} and M_{U2} involving continuous responses, the KLIC selected the true model approximately 85-90% of the time. For the comparisons of M_0 versus both M_{U1} and M_{U2} involving binary responses, the detection proportion dropped slightly, but the KLIC was

still able to identify the true model as the most ideal model approximately 65-80% of the time.

Table 6

Detection and Non-Detection Proportions of the Models Estimated in the Simulation

	Detection Proportion		Non-Detection Proportion	
	M_{U1}	M_{U2}	M_{O1}	M_{O2}
Case 1: Small sample binary outcome	0.72 (0.70, 0.74)	0.64 (0.61, 0.67)	0.35 (0.31, 0.39)	0.89 (0.86, 0.92)
Case 2: Large sample binary outcome	0.79 (0.77, 0.81)	0.72 (0.69, 0.75)	0.17 (0.14, 0.20)	0.87 (0.85, 0.89)
Case 3: Small sample continuous outcome	0.89 (0.86, 0.92)	0.87 (0.84, 0.90)	0.09 (0.08, 0.10)	0.41 (0.40, 0.42)
Case 4: Large sample continuous outcome	0.90 (0.89, 0.91)	0.84 (0.81, 0.87)	0.07 (0.06, 0.08)	0.49 (0.47, 0.51)

Note. KLIC is the Kullback-Leibler Information Criterion.
M_0 is the correct model. It has five predictors: 1 Type I TDC, 1 Type II TDC, and 1 Type III TDC.
M_{U1} is the underfit model that is missing an essential Type II TDC.
M_{U2} is the underfit model that is missing an essential Type III TDC.
M_{O1} is the overfit model with an additional Type II TDC.
M_{O2} is the overfit model with an additional Type III TDC.
Overall, the KLIC's detection proportion to detect the poor fit of underfit models was relatively high (greater than 0.50), and this was the case for both underfit models (M_{U1} and M_{U2}).
The KLIC's ability to distinguish an overfit model from the correct model was acceptable when the model was overfit with an additional Type II TDC (M_{O1}).
The KLIC performed poorly at selecting the correct model when an additional Type III TDC was included (M_{O2}).
Values in parentheses are the 95% standard binomial confidence limits.

The detection proportion of both comparisons--the true model versus the model without a necessary Type II TDC and the true model versus the model without a necessary Type III TDC--indicated that the KLIC was sensitive to distinguishing between good models and models that were missing critical information. For Case 4, which was the large-sample continuous data case, the KLIC did very well in selecting the true model over model M_{U1}, which was missing a necessary Type II TDC; it selected the true model over model M_{U1} 90% of the time. The section on "Model KLIC Averages" included a discussion about the Type II TDC having influence on responses across time. The omission of such an influential covariate sufficiently affected the model fit to the point that it was identified by the KLIC as an inferior model relative to the true model. Comparing these findings to the large-sample continuous data comparison of model M_{U1} to the true model, the detection proportion reduced by approximately 6%. As Type III TDCs only influence the response value at the current time period, it could be seen as having less of a feedback than Type II TDCs. The KLIC was able to correctly select the true model 84% of the time, but it was not as sensitive to the information lost from omitting a necessary Type III TDC as it was to the information lost from omitting a necessary Type II TDC.

For Cases 1 and 2, which were the binary cases, the KLIC's ability to correctly identify the true model as the most ideal model was still relatively good but not superior to the continuous data cases. Sample size seemed to affect the KLIC's performance as well--there was a small but noticeable jump in the detection proportion when the sample size grew from 500 to 2,500. The trends were the same as for those seen in the

continuous cases: the KLIC was able to more frequently select the true model when compared against model M_{U1}, which was missing a necessary Type II TDC.

The detection proportion of the comparisons of the true model and the underfit models tended to follow a general trend: comparisons involving the continuous outcome models tended to result in higher detection proportion than those involving binary outcome models. These results made sense because binary outcomes contain less precise information than do continuous outcomes in the sense that information have been dichotomized. Continuous data could be easily transformed into bins to create binary or categorical data, but the reverse involves more effort. For example, the transformation used in the Filipino Child Mortality (FCM) data analysis had to be postulated based on known mathematical properties of logarithms, as well as a conceptual understanding of the "story" behind the data. In this sense, transforming binary outcomes into continuous data interject additional information in the data, whereas information is lost when continuous data are dichotomized. It could be possible that the detection proportions of the continuous model comparisons were much higher than those of the binary model comparisons due to this information loss.

Discussion of non-detection proportion. Overall, the non-detection proportions were much higher than expected but detection proportion was also much higher than expected. The comparisons of the underfit models with the true model indicated that the KLIC was sensitive in its ability to recognize missing information; conversely, the comparisons of the overfit models with the true model suggested that the KLIC was not sensitive in its ability to recognize additional useless information--it tended to "reward" models that had more parameters.

The non-detection proportions of the second overfit model, M_{O2}, were much higher than expected for both the continuous and binary data analyses. As a model selection criterion, this implied that the KLIC was generous in selecting less parsimonious models. In particular, the non-detection proportions of the binary outcome models with the unnecessary Type III TDC, model M_{O2}, were extremely high, sufficiently high that the KLIC should not be used to select a candidate model in this type of data situation. The results of the simulation indicated that the KLIC preferred the overfit model over the true model almost 90% of the time when the outcome was binary. The non-detection proportions of the continuous outcome models with the same additional covariate were also relatively high. The overfit model was preferred over the true model approximately 40-50% of the time when the outcome was continuous. These results stipulated that the KLIC should be used with great caution when Type III TDCs are present in the data, especially when there is uncertainty about the inclusion of the covariate in the model estimation process.

The non-detection proportions of the comparisons of the model that was overfit with an additional Type II TDC, model M_{O1}, with the true model were also high but not as high as those of the comparisons involving model M_{O2}. The non-detection proportions of this comparison for the binary data analyses were much higher than those for the continuous data analyses. For binary outcome data with small sample sizes, the KLIC statistic preferred the overfit model a third of the time, on average. When the sample size tended to be larger, the KLIC preferred the overfit model a little less frequently. Again, these results were indicative of the KLIC's poor performance in identifying a good model when additional, unnecessary time dependent covariates are present in the data and the

outcome is binary. This was not necessarily the case for continuous data. When an unnecessary Type II TDC was included in estimating a continuous outcome model, the KLIC was able to correctly select the true model more than 90% of the time, regardless of sample size. Therefore, the KLIC's performance as a model selection criterion in the presence of additional Type II TDCs was superior for continuous data.

The general trend of the comparisons of the true model and the overfit models showed that the models with binary outcomes had higher non-detection proportions (i.e., the comparisons for Cases 1 and 2 had higher non-detection proportions than those for Cases 3 and 4). These results made sense, per the discussion in the "Discussion of Detection Proportion" section, regarding the loss or lack of information in dichotomized data, compared to the detail of information available in continuous data. It was possible that the non-detection proportions of the binary data analyses were much higher than those of the continuous data analyses due to the information lost in forcing data to be dichotomized into a binary response.

Comparisons of the Simulation and Data Analysis

The conclusions regarding the underfit models considered in the simulation could be applied to the analytic findings of the Filipino Child Mortality (FCM) data analysis: the detection proportion of the comparisons of the full model with the four candidate models should be relatively high. For the binary response analysis, the detection proportion of the comparisons could be expected to have been at least 60% but no higher than 80%, and the detection proportion of the comparisons of the continuous response analysis could be expected to have been approximately 85-90%, based on the values of detection proportion obtained in the simulation. The inclusion of a third underfit model in

the simulation--a model that was missing both an essential Type II and an essential Type III TDC--would have been helpful in the justification of the KLIC's selection of a model without BMI and survey round 2 as the fourth candidate model (M_4) in the binary response analysis.

One caveat of the continuous FCM data analysis is the transformation used to obtain the continuous response variable (Equation 15) which was proposed by Bhargava (1994) and Lai and Small (2007) in their studies. The original data set utilized a binary response denoting child morbidity (Bouis & Haddad, 1990), and it was transformed into a continuous response variable in order to estimate half of the models examined in the data analysis portion of this research. The log-transform in Equation 15 used to convert the binary response into a continuous response injected quite a bit of noise to the data, as the transformation took values bounded between 0 and 1 and transformed them into values on the real line, stretching from negative to positive infinity. The additional noise introduced by this transformation may have affected the findings of the continuous data analysis.

In contrast, the simulation data were generated under a much too controlled setting, in the sense that there was full control and manipulation of the data generation process, including the specification of the true model and the selection of the TDC feedback correlation coefficients. Values were selected based on trial and error so that they provided the most ideal conditions for the sake of time and computational resources, so it was likely that the generated data did not contain as much random noise as would real world data. In a nutshell, a simulation allows for the researcher to create a "story," then force the data to tell that story. In this sense, a simulation is flawed with some

artificiality, as it provides conditions that may or may not actually be encountered in the real world.

In the simulation, it was possible to distinguish the overfit models from the "correct model" because simulations allow the researcher to specify the desired conditions, then generate data to follow those conditions. The KLIC tended to prefer overfit models in the simulation; meanwhile, it selected the full model as the most ideal model in the FCM analysis. As suggested by previous studies (Bhargava, 1994; Bouis & Haddad, 1990; Lai & Small, 2007), it was possible that BMI was, indeed, a meaningless predictor of child morbidity in the Philippines. It was conceivable that the "true" model for the FCM data was the model without BMI, but the KLIC's inclination to select overfit models could have resulted in the full model being favored in both the binary and continuous outcome analyses. This theory supports the marginally significant *p*-values provided by the chi-squared comparison of the full model and the model without BMI for both the binary and continuous outcome analyses (approximately 0.01 in both cases). Evidently, some fine-tuning of the KLIC would be necessary prior to its regular use as an information-based model selection method.

CHAPTER V

CONCLUSIONS

In this research, a novel use of the Kullback-Leibler divergence principle was applied to conceive the Kullback-Leibler Information Criterion (KLIC) as a model selection criterion for models estimated using Generalized Method of Moments (GMM) when time-dependent covariates (TDCs) were present in the data. A simulation study, in conjunction with the analysis of real data, was used to understand the performance of the KLIC under multiple scenarios. Overall, the performance of the KLIC was better than expected in the simulation study when the response was normally distributed, except when an extraneous Type III TDC was included in the model. This was not the case for binary response data: the simulation showed that the KLIC was able to correctly select the true model against an underfit model, but it performed poorly when additional, unnecessary TDCs were included in the model. It frequently selected the overfit model over the true model, and this was especially the case when the extraneous predictor was a Type III TDC.

This chapter begins with a brief recap of the results and findings of this research, tying in the findings from the data analysis with those of the simulation. The subsequent section presents the limitations of this research, including a discussion about future research to improve the proposed statistic's ability to better serve its purpose as an information criterion to identify the most ideal model from a collection of candidates.

The final section in Chapter V considers applications and extensions of this research for the applied researcher.

Overall Findings

Although the detection proportion of the comparisons of the true model with the underfit models were high for all four simulation cases, the non-detection proportions for three of the four data situations were also very high. Based on the non-detection proportions, the KLIC was able to successfully select the correct model at least 90% of the time only when the response variable was continuous and the true model was compared to a model that included an additional, unnecessary Type II TDC. This conclusion did not carry over when the extraneous covariate was swapped to a Type III TDC. It also did not apply to the binary data analysis. Some modification to the KLIC would be necessary for it to be used as a model selection standard in binary data situations.

Based on the results of the simulation, the KLIC's performance as a model selection criterion showed that it achieved good detection proportion in identifying a good model when key variables were left out, but it had high non-detection rates in selecting the true model when extraneous variables were introduced to candidate models--i.e., the KLIC tended to favor more complex models when there was a lot of noise in the data. This was especially the case for binary response data in the presence of extraneous Type III TDCs, with non-detection proportions as high as 87-89%, although continuous response data with Type III TDCs also had very high non-detection proportions.

Limitations and Suggestions for Further Research

This research considered the use of the Kullback-Leibler Information Criterion (KLIC) as a model selection criterion for models estimated using Generalized Method of Moments (GMM) in the presence of time-dependent covariates (TDCs). In the analysis of the Filipino Child Mortality (FCM) data, two scenarios were examined--one that utilized the original binary response, indicating the child's morbidity, and a continuous response, which used a transformation to denote a continuous measure of morbidity. To augment the findings of the data analysis and provide a more extensive view of the performance of the KLIC, a simulation study was also conducted. The simulation considered two different response types—binary and continuous—and small and large sample size scenarios, and it varied the number and types of TDCs included or excluded in the model, comparing the results to the performance of the KLIC for the "true" model.

Details of the findings and conclusions from the analysis and simulation were included in Chapter IV, and the following sections describe the limitations of those findings. The discussions of the limitations include potential fixes, as well as ideas for future research, including the impact of individual limitations on the current research and the potential improvements made possible by further explorations.

Penalty for Model Complexity

The most evident limitation of the current research was the lack of a penalty term in the Kullback-Leibler Information Criterion (KLIC), similar to those exercised by Akaike's Information Criterion and the Bayesian Information Criterion. Adding a penalty for less parsimonious models could be helpful in adjusting the KLIC's selection of overfit models, as was the case in the simulation, especially for binary response data, although it

could also benefit the KLIC's performance for continuous outcome data. As the results of the simulation showed, the KLIC tended to select overfit data over underfit data, and the non-detection proportion for the selection of the overfit data over the true model was as high as almost 90% for the binary data and as high as almost 50% for the continuous data. Current results suggested that, without a thorough knowledge of the story behind the data, the KLIC could not be recommended as a model selection criterion for GMM modeling of binary outcome data when TDCs were present, as it would almost always select the full model, regardless of its inclusion of unnecessary parameters. In some ways, its behavioral property was similar to that of the coefficient of determination (R^2), to the degree that R^2 tends to grow closer to 1 as additional variables are included in the model, regardless of its true relationship with the response. As the adjusted coefficient of determination (adjusted R^2) is recommended for models with multiple predictors to control for its inflation with additional parameters, an adjusted KLIC is necessary to account for the number of time-dependent predictors included in the model.

Further research would be necessary to formulate the exact penalty term of the KLIC and what information it should include, but one suggestion is that simpler is better. For example, the AIC's penalty is +2k, where k is the number of parameters estimated, and the BIC's penalty is a function of the number of parameters in the model and the number of observations in the data. The penalty terms of these existing information criteria were chosen based on the relevance of these two pieces of information in the estimation process itself--for example, the penalty for the BIC is +2k*ln(n), where n denotes the number of observations in the data, i.e., the sample size. The BIC is a large-sample asymptotic approximation to the full Bayesian model comparison (Busemeyer &

Diederich, 2015), so the inclusion of information about both the sample size and the number of model parameters in its penalty term was a sensible mathematical choice. Analogously, the KLIC's penalty term should be a function of the number of moment conditions associated with each TDC, as well as some information about the TDC type, which are two critical pieces of information used in the GMM estimation process. Therefore, a reasonable starting point for future research involving the penalty term for the KLIC may be, for example, +2k*j, where k would be the number of parameters in the model and j would be the number of moment conditions used in the GMM estimation.

Small-Sample Correction

Looking at the overall trend in the detection and non-detection proportions of the comparisons made in the simulation, the KLIC's ability to select the true model over the underfit and overfit models was almost always improved with an increase in the sample size. This was the case across the board except for two continuous outcome scenarios: once when the model was overfit with an additional Type III TDC (model M_{O2}) and again when an important Type III TDC was deliberately left out (model M_{U2}). However, these discrepancies were minor and have been left aside in this discussion. For the majority of the scenarios, the inclusion of a small-sample correction to the KLIC may be beneficial in applied research, as real-world data could oftentimes be limited in their sample sizes.

The corrected AIC, or AICc, is a small-sample correction of the AIC, which was a simple modification of the AIC's penalty term from + 2k to + 2k + (2k (k+1) / (n-k-1)), essentially incorporating information about the sample size to the AIC for finite samples. The discrepancy between the AIC and AICc diminish with increasing sample size (Anderson & Burnham, 2002; Burnham & Anderson, 2004; Hurvich & Tsai, 1995); as

such, Burnham and Anderson (2004) encouraged the regular use of the AICc in every analysis, regardless of the sample size. A similar small-sample correction to the KLIC could be the inclusion of a term that incorporates the number of parameters, the number (or types) of moment conditions used in the model estimation, and the number of observations in the data. Conscious effort should be taken to ensure that the limit of the correction term approach 1 as n tends to ∞ so that the recommendation ensue that the small-sample adjusted KLIC, or KLICc, be standardly used to select candidate GMM models in the presence of TDCs.

Performance Under Non-Binary and Non-Gaussian Responses

The current research considered only two types of responses: one that followed a binominal distribution and another that followed the Gaussian normal distribution. These were only two of the many response distributions encountered in applied research, and the performance of the KLIC should be investigated for other response distributions. For example, survival time of the sick children whose data were collected for the FCM data could have been modeled by the exponential distribution, or the time between the last day of illness and mortality could have been modeled by the Gamma distribution. Even the original response, which was count data representing the number of sick days, could have been modeled--as is--as Poisson data. The performance of the KLIC should be assessed for Non-Gaussian responses such as these, to provide a more comprehensive look at the KLIC's ability to select among candidate models when the response is not binary or normally distributed.

One caveat to this discussion would be that the KLIC may be quite robust in selecting among candidate models estimated using GMM, as the estimation process itself

did not require knowledge about the response distribution, but rather, used only moment conditions with zero expectation. It is possible that the comparisons examined in the simulation study under alternate response distributions may yield similar results as have already been shown in this research. However, it would be advisable to test the performance of the KLIC for different data types prior to its use as the primary, or only, model selection criterion for models estimated using GMM in the presence of time-dependent covariates for non-binary and non-Gaussian data.

Performance in the Presence of Type IV Time-Dependent Covariates

The current research was limited in its inclusion of all TDC types--it excluded Type IV TDCs from all of the analyses. The reason for this omission was threefold: (a) the FCM data did not include apparent Type IV TDCs, so it was not included in the data analyses; (b) the simulation data were intended to mirror, to some extent, the structure of the FCM data, and hence, TDCs of Types I, II, and III were included in the simulation data; and (c) TDCs of Type IV are less common in the real world, and the feedback process is a little more challenging to understand, as well as program. On the contrary, the feedback process of this type of TDC is the reverse of Type II TDCs; consequently, the assumption was made that the performance of the KLIC in model comparisons involving Type IV TDCs should mimic its performance in comparing models involving Type II TDCs. This may or may not necessarily be the case.

The natural next step following this research is to repeat the simulation study with the inclusion of a Type IV TDC; namely, to extend the true model to one that also included a Type IV TDC and examine changes in model selection using the KLIC when

this type of time-varying covariate is added to the mix. To do this, two additional models should be examined: model M_{U3}, which would be a model that was missing the essential Type IV TDC, and model M_{O3}, which would be a model with an additional, unnecessary Type IV TDC. The use of the KLIC as a model selection criterion for the GMM modeling of data with TDCs of Type IV would not be recommended until such a simulation study has been conducted and the KLIC's performance under such data scenario assessed.

Time-Dependent Covariate Feedback Loop Correlations

The random data generation process for the simulation was set up in such a way that it involved "trial-by-error." A crucial assumption behind this research was that this process was set up correctly to generate the desired data as specified so that (a) the true model was, indeed, the true model for these data and (b) the time-dependent covariates generated for the analysis were actually specified and generated to reflect the appropriate feedback loops with reasonable and realistic feedback correlation coefficients.

There currently has been no known software that randomly generates time-dependent covariates, and this research involved the fabrication of a program that did so. The values used to generate the feedback process for the time-dependent covariates were specified in this research after multiple iterations of trial-and-error. The final correlations used to generate the feedback process were: $\rho_{xy} = 0.25$ for the Type II TDCs and $\rho_{yy} = 0.25$ for the Type III TDCs.

Different values of ρ_{xy} and ρ_{yy} provided different response values, and there were countless possibilities of these correlations, which provided numerous outcomes. A more systematic method to selecting these correlations should be pursued in a future research,

such as determining thresholds of ρ_{xy} and ρ_{yy} for strong or weak feedback loops. The performance of the KLIC should be tested for these thresholds to determine general trends or changes based on the strength of the feedback correlation. The KLIC's ability to select among candidate models is expected to improve with greater feedback, as the number and types of moment conditions used in the computation of the KLIC depended on TDC type. The weaker this correlation, the weaker the time-dependency of the feedback; this would mean the TDC's distinction from a time-independent covariate is negligible and other methods of model selection, such as the chi-squared test used in the FCM analysis, would suffice in the model selection process.

The specification of TDCs and the magnitude of the strength of TDC feedback loops were not the main focus of this research. The main focus was on the performance of the KLIC statistic in selecting the most ideal model among multiple candidates. The discussion regarding various strengths in the TDC feedback loops and their effects on the KLIC's performance has been set aside for future research.

Binary Time-Dependent Covariates

Three of the five predictors in the Filipino Child Mortality (FCM) analysis were binary, but only continuous predictors were included in the simulation. Concerns about separation, GMM algorithm convergence, and other issues that occasionally arise due to the inclusion of binary predictors were some of the reasons for this decision. As the simulation analyses for both the binary and continuous outcome data included only continuous predictors, the results and conclusions based on the simulation only apply to data with continuous predictors. In order to better understand the behavior of the KLIC

under data scenarios involving binary time-dependent covariates, an extension of the research presented in Chapter IV would be necessary.

One challenge foreseen in simulating binary TDCs is the feedback loop inherent in the time dependency of these predictors, as discussed in the preceding section. There was some challenge in writing a program that incorporated the feedback--namely, ρ_{xy} and ρ_{yy} --from the time-varying predictor to the continuous and binary outcome. In particular, for the binary data generation, feedback correlations were used to generate a set of continuous predictors, and a combination of these correlations was used to generate a binary outcome, or 0s and 1s. There is some information loss in the coercion of a number to a binary, especially when rich information is available in the set of predictors. To replicate this for binary TDCs, the process would involve using binary information to predict a binary outcome, and this would introduce even more information loss throughout the process. Further research would be necessary to design a robust method of randomly generating binary TDCs, especially when the outcome is also binary.

Further Discussions

To gain further insight into the performance of the KLIC or its ability to select the most ideal model for the data, a resampling method could be considered to test the goodness-of-fit of the model on subsamples of the data, rather than looking at it for the overall data, per case. Another option could be to use cross validation to check that the estimated models actually make sense; basically, this would be model optimization using cross validation. This would be one way to assess the KLIC's ability to select among a collection of underfit and overfit models, and it can be more informative than looking at the non-detection proportions of the tests involving the overfit models.

On the flipside, concerns regarding the under-fitting versus overfitting dilemmas have to do with how a model is applied to new data, which was not the main focus of this research. The goal of this research was not to estimate the "best" model for the sake of prediction; it was to investigate the ability of the KLIC to select the most ideal model for the data.

Applications for Applied Research

This research filled a large niche in the current gap of knowledge by introducing an alternative method of model selection for models estimated using Generalized Method of Moments (GMM) in the presence of time-dependent covariates (TDCs). Any applied researcher who handles longitudinal data should be mindful of the presence of TDCs in their data, as the feedback introduced by these predictors should be accounted for in model estimation. Lai and Small (2007) showed that 2-Step GMM (2SGMM) was superior to GEE and other estimation methods in its estimation of model parameters when TDCs were present in the data. The renowned statement, "all models are wrong; some models are useful," (Box & Draper, 1987) captures the significance of model selection in any applied research. Numerous models could be estimated from a given set of data; many are poor approximations, but a few are reasonable summaries of the full information underlying the data.

The Kullback-Leibler Information Criterion (KLIC) provided an alternative to the current method; further, it accounted for the feedback that was inherent in most longitudinal data, which the chi-squared method currently does not take into consideration. Ignoring the feedback loop introduced by time-varying covariates and relying on the regular use of the Generalized Estimating Equations (GEE) for the analysis

of longitudinal data could compromise model parameter consistency, efficiency, and bias, resulting in misleading inferences (Diggle et al., 2002; Fitzmaurice, 1995; Fitzmaurice et al., 2011; Pepe & Anderson, 1994). The applied researcher has the responsibility to distinguish the suitable models from poor estimates of the full data--this obligation drove the need for a reliable method to make this important distinction. Replacing the former practice with the routine use of GMM to properly account for feedback in the data is highly encouraged. Those who do so would find the use of the KLIC beneficial in their daily work, as they would need the right tool to select an appropriate model among a collection of candidate GMM models, especially when there are time-varying predictors in the data.

Final Remarks

This research showed that the information criterion derived from the Kullback-Leibler divergence principle could be used to select among candidate Generalized Method of Moments (GMM) models when time-varying predictors were present in the data. A simulation and the analyses of real data were used to examine the performance of the Kullback-Leibler Information Criterion (KLIC) under various data settings. Overall, the KLIC performed well as a model selection method for normally distributed response data, and its ability to identify an appropriate candidate model in the presence of Type III time-dependent covariates (TDCs) could be improved after some refinement. For binary response data, the KLIC showed good potential in its ability to identify against underfit models but functioned poorly against overfit models, almost always selecting the overfit model over the true model. The modifications provided in the Limitations and Suggestions for Further Research section were recommended in fine-tuning the KLIC's

ability to more accurately select the best candidate model were it to become the standard information criterion in selecting among candidate GMM models in the presence of TDCs.

REFERENCES

Agresti, A. (1990). *Categorical data analysis*. Hoboken, NJ: John Wiley & Sons, Inc.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proceedings of the Second International Symposium on Information Theory*, 267-281.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723.

Altonji, J. G., & Segal, L. M. (1996). Small-sample bias in GMM estimation of covariance structures. *Journal of Business & Economics Statistics, 14*(3), 353-366.

Anderson, D. R., & Burnham, K. P. (2002). Avoiding pitfalls when using information-theoretic methods. *The Journal of Wildlife Management, 66*(3), 912-918.

Andrews, D. W. K. (1999). Consistent moment selection procedures for generalized method of moments estimation. *Econometrica, 67*(3), 543-564.

Bain, L. J., & Engelhardt, M. (1992). *Introduction to probability and mathematical statistics* (2nd ed.). Belmont, CA: Duxbury Press.

Bhargava, A. (1994). Modeling the health of Filipino children. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 157*(3), 417-432.

Bouis, H. E., & Haddad, L. J. (1990). Effects of agricultural commercialization on land tenure, household resource allocation, and nutrition in the Philippines (Research Report 79). Washington, DC: International Food Policy Research Institution.

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. New York, NY: John Wiley & Sons, Inc.

Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics, 53*(2), 603-618.

Burnham, K. P., & Anderson, D. R. (1998). Model selection and inference: A practical information-theoretic approach. New York, NY: Springer.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research, 33*(2), 261-304.

Busemeyer, J. R., & Diederich, A. (2015). *Neuroeconomics*. Retrieved February 4, 2018, from https://www.sciencedirect.com/topics/medicine-and-dentistry/akaike-information-criterion.

Cavanaugh, J. E., & Neath, A. A. (1999). Generalizing the derivation of the Schwarz information criterion. *Communications in Statistics: Theory and Methods, 28*(1), 49-66.

Chaurasia, A., & Harel, O. (2013). Model selection rates of information based criteria. *Electronic Journal of Statistics, 7*, 2762-2793.

Chaussé, P. (2010). Computing generalized method of moments and generalized empirical likelihood with *R*. *Journal of Statistical Software, 34*(11), 1-35.

Csiszár, I. (1975). *I*-divergence geometry of probability distributions and minimization problems. *The Annals of Probability, 3*(1), 146-158.

Cui, J. (2007). QIC program and model selection in GEE analyses. *The Stata Journal, 7*(2), 209-220.

Diggle, P. J., Heagerty, P., Liang, K-Y., & Zeger, S. L. (2002). *Analysis of longitudinal data*. New York, NY: Oxford University Press Inc.

Dobson, A. J., & Barnett, A. G. (2008). *An introduction to generalized linear models* (3rd ed). Boca Raton, FL: Chapman and Hall, CRC Press.

Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics, 51*, 309-317.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis* (2nd ed). Hoboken, NJ: John Wiley & Sons, Inc.

Hall, A. R. (1999). Hypothesis testing in models estimated by GMM. *Generalized Method of Moments Estimations*. Cambridge, UK: Cambridge University Press. 96-127.

Hall, A. R. (2005). *Generalized method of moments, advanced texts in econometrics*. New York, NY: Oxford University Press, Inc.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica, 50*(4), 1029-1054.

Hansen, L. P. (2007). Generalized method of moments estimation. *Department of Economics, University of Chicago Lecture*: Chicago, IL.

Hansen, L. P., Heaton, J., & Yaron, A. (1996). Finite sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics, 14*(3), 262-280.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, NJ: Wiley & Sons, Inc.

Hilbe, J. M. (2009). *Logistic regression models*. Boca Raton, FL: Chapman and Hall/CRC.

Hin, L. Y., & Wang, Y. G. (2009). Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine, 28,* 642-658.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). Hoboken, NJ: Wiley & Sons, Inc.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed). Hoboken, NJ: Wiley & Sons, Inc.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*(2), 297-307.

Hurvich, C. M., & Tsai, C.-L. (1995). Model selection for extended quasi-likelihood models in small samples. *International Biometric Society, 51*(3), 1077-1084.

Imbens, G. W. (1997). One-step estimators for over-identified generalized method of moments models. *Review of Economic Studies 64*, 359-383.

Kitamura, Y., & Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrics 65*(4), 861-874.

Kleiber, C., & Zeileis, A. (2008). *Applied econometrics with R*. New York, NY: Springer.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods and Research, 33*(2), 188-229.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79-86.

Lai, T. L., & Small, D. (2007). Marginal regression analysis of longitudinal data with time-dependent covariates: a generalized method-of-moments approach. *Journal of the Royal Statistical Society, Series B, 69*(1), 79-99.

Lalonde, T. L., Wilson, J. R., & Yin, J. (2014). GMM logistic regression models for longitudinal data with time-dependent covariates and extended classifications. *Statistics in Medicine, 33*(27), 4756-4769.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13-22.

Liang, K. Y., Zeger, S. L., & Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, 54*(1), 3-40.

Lin, H. Y., & Myers, L. (2006). Power and type I error rates of goodness-of-fit statistics for binomial generalized estimating equations (GEE) models. *Computational Statistics & Data Analysis, 50*, 3232-3448.

Mátyás, L. (1999). *Generalized method of moments estimation*. Cambridge, England: Cambridge University Press.

McCullagh, P., & Nelder, J. A. (1983). Quasi-likelihood functions. *Annals of Statistics, 11*, 59-67.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. Boca Raton, FL: Chapman and Hall, CRC Press.

Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Mendenhall, R., Scheaffer, R. L., & Wackerly, D. D. (1981). *Mathematical statistics with applications*. Belmont, CA: Duxbury Press.

Neuhaus, J. M., & Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics, 54*(2), 638-645.

Newey, W. K. (1985). Generalized method of moments specification testing. *Journal of Econometrics, 29*, 229-256.

Nielsen, H. B. (2005). Generalized method of moments estimation. *Econometrics 2 Lecture, Department of Economics, University of Copenhagen*: Copenhagen, Denmark.

Pan, W. (2001a). Akaike's information criterion in generalized estimating equations. *Biometrics,* 57, 120-125.

Pan, W. (2001b). Sample size and power calculations with correlated binary data. *Division of Biostatistics, School of Public Health Periodical*: Minnesota, MN.

Pepe, M. S., & Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics: Simulation and Computation, 23*(4), 939-951.

Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology, 53*(5), 793-808.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics, 9*(4), 705-724.

Rodríguez, C. C. (2005). The ABC of model selection: AIC, BIC, and the New CIC. *Bayesian Inference and Maximum Entropy Methods in Science and Engineering, CP803*, 80-87.

Sayyareh, A., Obeidi, R., & Bar-Hen, A. (2011). Empirical comparison between some model selection criteria. *Communications in Statistics: Simulation and Computation, 40*, 72-86.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461-464.

Shane, M. N. (2013). Estimating parameters for binary data with time-dependent covariates using the generalized method of moments. In JSM Proceedings, Biometrics Section. Alexandria, VA: American Statistical Association. 2338-2347.

Sommer, A., Katz, J., & Tarwotjo, I. (1984). Increased risk of respiratory disease and diarrhea in children with preexisting mild vitamin A deficiency. *American Journal of Clinical Nutrition, 40*, 1090-1095.

Sommer, A., Tarwotjo, I., Hussaini, G., & Susanto, D. (1983). Increased mortality in children with mild vitamin A deficiency. *Lancet, 2*, 585-588.

Sommer, A., Tarwotjo, I., & Katz, J. (1987). Increased risk of xerophthalmia following diarrhea and respiratory disease. *American Journal of Clinical Nutrition, 45*, 977-980.

Stigler, S. M. (2008). Karl Pearson's theoretical errors and the advances they inspired. *Statistical Science, 23*(2), 261-271.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics A (Theory and Methods)*, A7: 13-26.

Wackerly, D. D., Mendenhall, W. III, & Scheaffer, R. L. (2008). *Mathematical statistics with applications* (7th ed.). Belmont, CA: Duxbury Press.

Wang, Y., & Liu, Q. (2006). Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of stock-recruitment relationships. *Fisheries Research, 77,* 220-225.

Ward, E. J. (2008). A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. *Ecological Modeling, 211*, 1-10.

Weakliem, D. L. (1999). A critique of the bayesian information criterion for model selection. *Sociological Methods and Research, 27*(3), 359-397.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika, 61*(3), 439-447.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica, 50*(1), 1-25.

Zeger, S. L., & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42, 121-130.

Zeger, S. L., & Liang, K. Y. (1991). Feedback models for discrete and continuous outcomes. *Statistica Sinica,* 1, 51-64.

Zeger, S. L., & Liang, K-Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine, 11*, 1825-1839.

Zeger, S. L., Liang, K. Y., & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation Approach. *Biometrics 44*(4), 1049-1060.

Zivot, E. (2015). Generalized method of moments. *Econometrics 583 Lecture at the University of Washington*: Seattle, WA.

Zsohar, P. (2012). Short introduction to the generalized method of moments. *Hungarian Statistical Review, 16,* 150-170.

APPENDIX A

CHAPTER II DEFINITIONS

CHAPTER II DEFINITIONS

Consistency:

The idea of *consistency* is related to the value of the parameter estimate as n approaches ∞. A *consistent estimator* is one that approaches the true value of the parameter as the sample size increases (Bain & Engelhardt, 1992).

Let $\hat{\theta}$ be an estimator for θ. Then, $\hat{\theta}$ is a *consistent estimator* of θ if:

$$\forall\ \varepsilon > 0, \lim_{n\to\infty} P\left[\left|\hat{\theta} - \theta\right| > \varepsilon\right] = 0\,.$$

Efficiency:

The *relative efficiency* of an unbiased estimator $\hat{\theta}$ of θ to another unbiased estimator, θ*, is:

$$r.e.\left(\hat{\theta}, \theta^*\right) = \frac{Var\left(\theta^*\right)}{Var\left(\hat{\theta}\right)}.$$

An unbiased estimator θ* of θ is *efficient* if $r.e.\left(\hat{\theta}, \theta^*\right) \le 1$ for all unbiased estimators of θ (Bain & Engelhardt, 1992). An *efficient estimator* is one that has uniformly minimum variance as the sample size increases (i.e., asymptotic minimum variance).

APPENDIX B

CHAPTER IV KULLBACK-LEIBLER INFORMATION CRITERION STATISTICS OF ALL 31 MODELS FOR THE BINARY OUTCOME ANALYSIS OF THE FILIPINO CHILD MORTALITY DATA

CHAPTER IV KULLBACK-LEIBLER INFORMATION CRITERION STATISTICS OF ALL 31 MODELS FOR THE BINARY OUTCOME ANALYSIS OF THE FILIPINO CHILD MORTALITY DATA

Table 7

Kullback-Leibler Information Criterion of All 31 Models for the Binary Data Analysis

	Kullback-Leibler Information Criterion (KLIC)	Variables included
Most ideal model (M1)	148.7	Age, gender, BMI, survey round 2, survey round 3 (i.e., the full model)
2nd candidate model (M2)	167.9	Age, gender, survey round 2, survey round 3
3rd candidate model (M3)	202.3	Age, gender, survey round 3, BMI
4th candidate model (M4)	251.1	Age, gender, survey round 3
5th candidate model (M5)	311.4	Age, BMI, survey round 2, survey round 3
6th candidate model (M6)	315.2	BMI, age, gender, survey round 2
7th candidate model (M7)	331.1	Age, survey round 3, survey round 2
8th candidate model (M8)	335.7	Age, gender, survey round 2
9th candidate model (M9)	347.8	BMI, age, gender
10th candidate model (M10)	349.4	BMI, gender, survey round 2, survey round 3
11th candidate model (M11)	365.9	BMI, age, survey round 3
12th candidate model (M12)	368.1	BMI, age, survey round 2
13th candidate model (M13)	375.3	gender, survey round 2, survey round 3

Table 7 (continued)

	Kullback-Leibler Information Criterion (KLIC)	Variables included
14th candidate model (M14)	379.6	BMI, gender, survey round 3
15th candidate model (M15)	381.3	BMI, gender, survey round 2
16th candidate model (M16)	389.2	BMI, survey round 2, survey round 3
17th candidate model (M17)	404.8	Age and gender
18th candidate model (M18)	411.5	Age and survey round 3
19th candidate model (M19)	413.1	Age and survey round 2
20th candidate model (M20)	420.0	Gender and survey round 3
21st candidate model (M21)	422.2	Gender and survey round 2
22nd candidate model (M22)	431.7	BMI and age
23rd candidate model (M23)	440.6	Survey round 2 and survey round 3
24th candidate model (M24)	441.9	BMI and gender
25th candidate model (M25)	472.4	Age
26th candidate model (M26)	477.8	BMI and survey round 3
27th candidate model (M27)	479.3	BMI and survey round 2
28th candidate model (M28)	498.2	Gender
29th candidate model (M29)	519.6	Survey round 3
30th candidate model (M30)	521.2	Survey round 2
31st candidate model (M31)	545.7	BMI

Note. BMI = body mass index

APPENDIX C

CHAPTER IV KULLBACK-LEIBLER INFORMATION CRITERION STATISTICS OF ALL 31 MODELS FOR THE CONTINUOUS OUTCOME ANALYSIS OF THE FILIPINO CHILD MORTALITY DATA

CHAPTER IV KULLBACK-LEIBLER INFORMATION CRITERION STATISTICS OF ALL 31 MODELS FOR THE CONTINUOUS OUTCOME ANALYSIS OF THE FILIPINO CHILD MORTALITY DATA

Table 8

Kullback-Leibler Information Criterion of All 31 Models for the Continuous Data Analysis

	Kullback-Leibler Information Criterion (KLIC)	Variables included
Most ideal model (M1)	287.9	Age, gender, BMI, survey round 2, survey round 3 (i.e., the full model)
2nd candidate model (M2)	293.5	Age, gender, survey round 2, survey round 3
3rd candidate model (M3)	320.3	BMI, age, gender, survey round 3
4th candidate model (M4)	324.1	BMI, age, gender, survey round 2
5th candidate model (M5)	376.6	BMI, age, survey round 2, survey round 3
6th candidate model (M6)	381.2	BMI, age, gender, survey round 2
7th candidate model (M7)	390.7	Age, gender, survey round 2
8th candidate model (M8)	408.2	Age, survey round 3, survey round 2
9th candidate model (M9)	421.8	BMI, age, gender
10th candidate model (M10)	433.1	BMI, age, survey round 3
11th candidate model (M11)	437.3	BMI, age, survey round 2

Table 8 (continued)

	Kullback-Leibler Information Criterion (KLIC)	Variables included
12th candidate model (M12)	450.8	BMI, gender, survey round 2, survey round 3
13th candidate model (M13)	455.4	Age and gender
14th candidate model (M14)	474.7	Age and survey round 3
15th candidate model (M15)	478.1	Age and survey round 2
16th candidate model (M16)	491.1	Gender, survey round 2, survey round 3
17th candidate model (M17)	528.3	BMI, gender, survey round 3
18th candidate model (M18)	531.7	BMI, gender, survey round 2
19th candidate model (M19)	TDC6.3	BMI and age
20th candidate model (M20)	615.2	BMI, survey round 2, survey round 3
21st candidate model (M21)	638.6	Gender and survey round 3
22nd candidate model (M22)	641.1	Gender and survey round 2
23rd candidate model (M23)	653.2	Age
24th candidate model (M24)	675.2	Survey round 2 and survey round 3
25th candidate model (M25)	698.4	BMI and gender
26th candidate model (M26)	707.7	Gender
27th candidate model (M27)	746.1	BMI and survey round 3
28th candidate model (M28)	749.0	BMI and survey round 2
29th candidate model (M29)	761.8	Survey round 3

Table 8 (continued)

	Kullback-Leibler Information Criterion (KLIC)	Variables included
30th candidate model (M30)	765.2	Survey round 2
31st candidate model (M31)	793.3	BMI

Note. BMI = body mass index

APPENDIX D

BOOTSTRAPPED KULLBACK-LEIBLER INFORMATION CRITERION FOR THE FIVE MODELS ESTIMATED IN THE SIMULATION

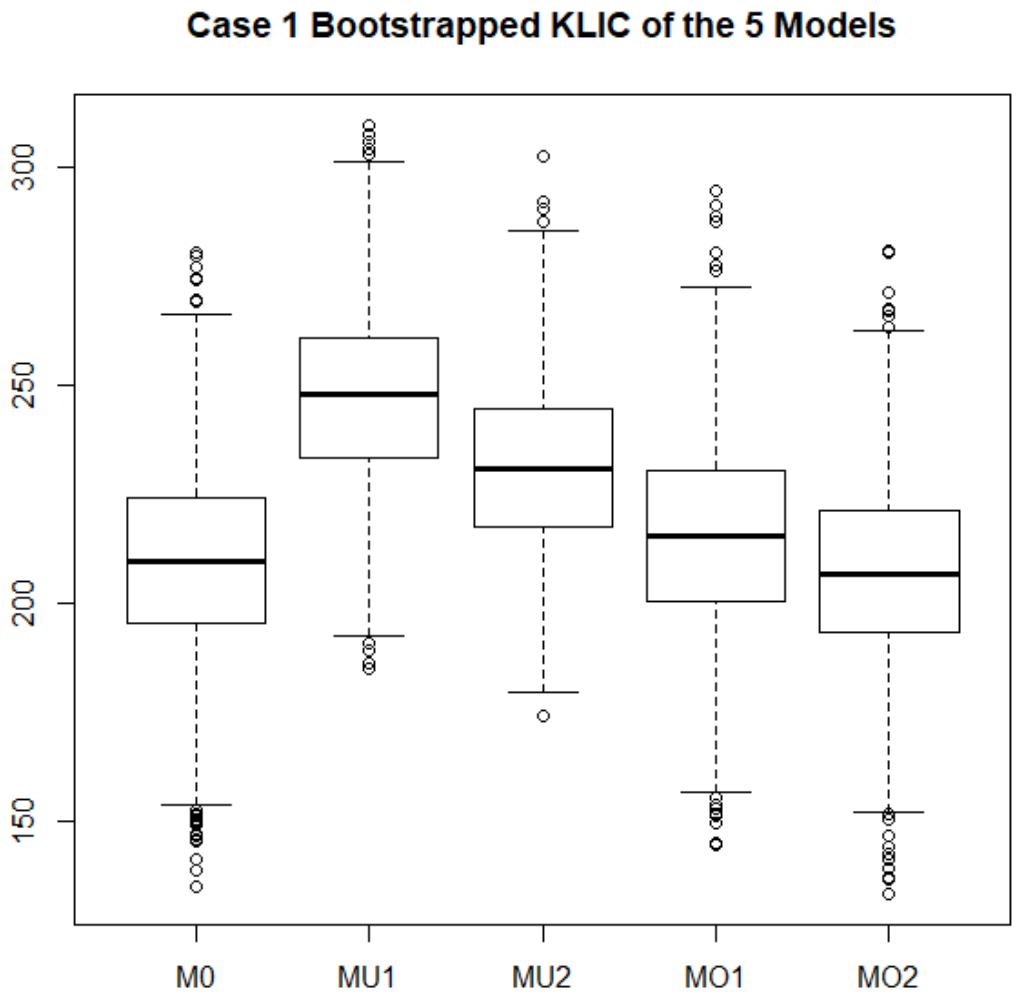

Figure 1. Kullback-Leibler Information Criterion of the 5 Models Estimated in the Simulation for the Small-Sample Binary Outcome Case (Case 1).

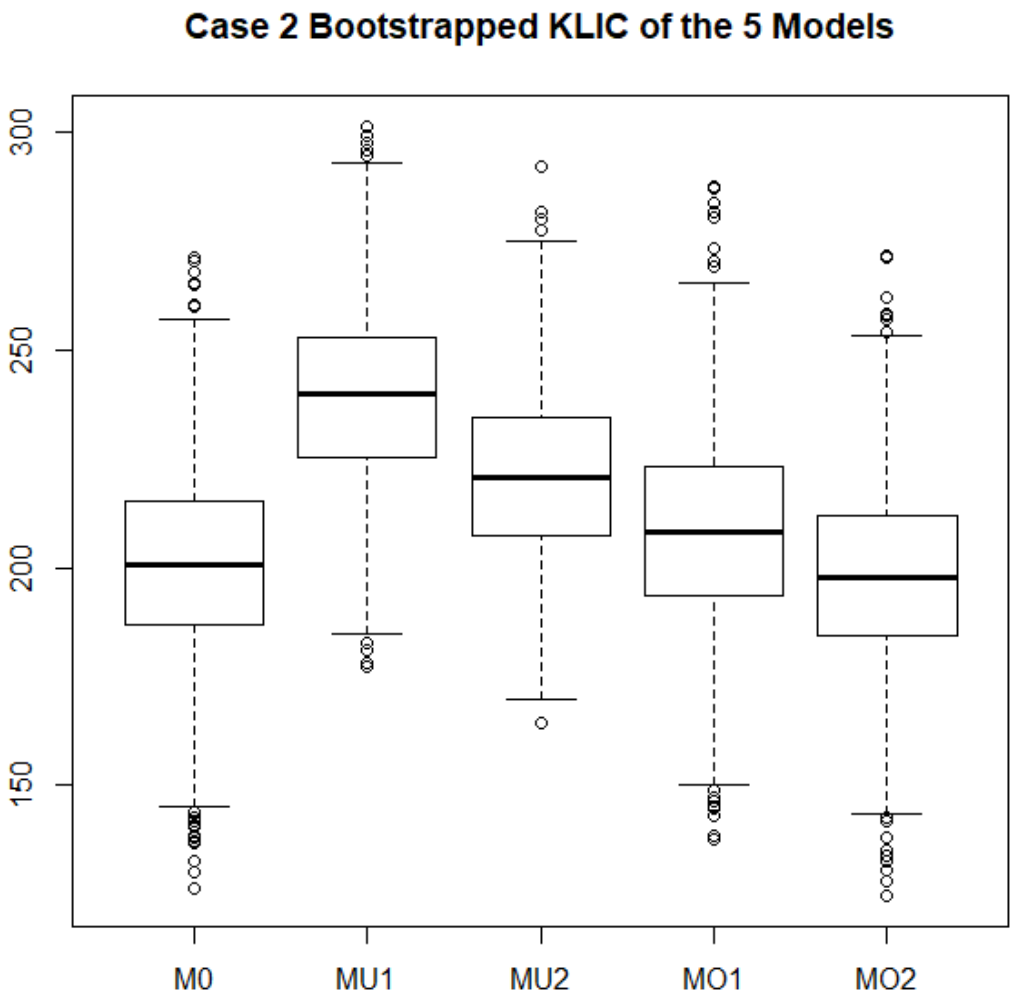

Figure 2. Kullback-Leibler Information Criterion of the 5 Models Estimated in the Simulation for the Large-Sample Binary Outcome Case (Case 2).

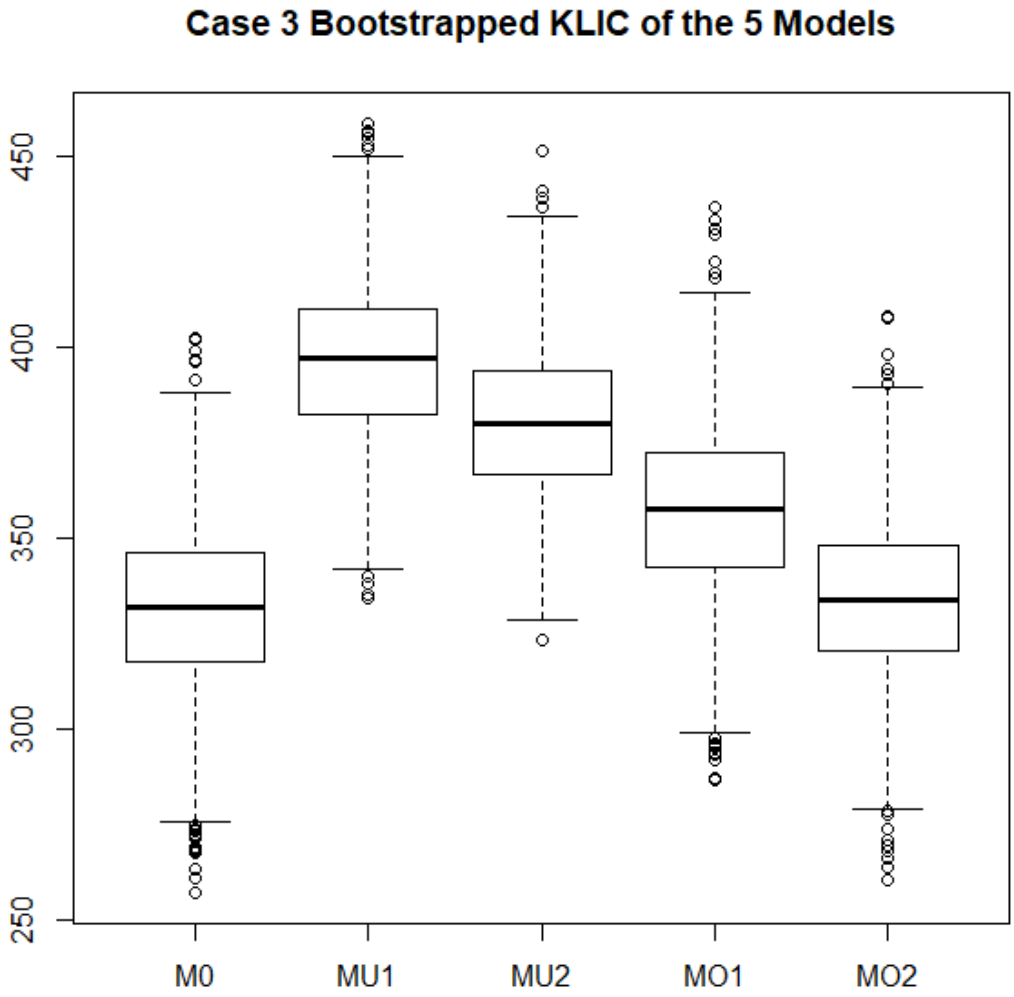

Figure 3. Kullback-Leibler Information Criterion of the 5 Models Estimated in the Simulation for the Small-Sample Continuous Outcome Case (Case 3).

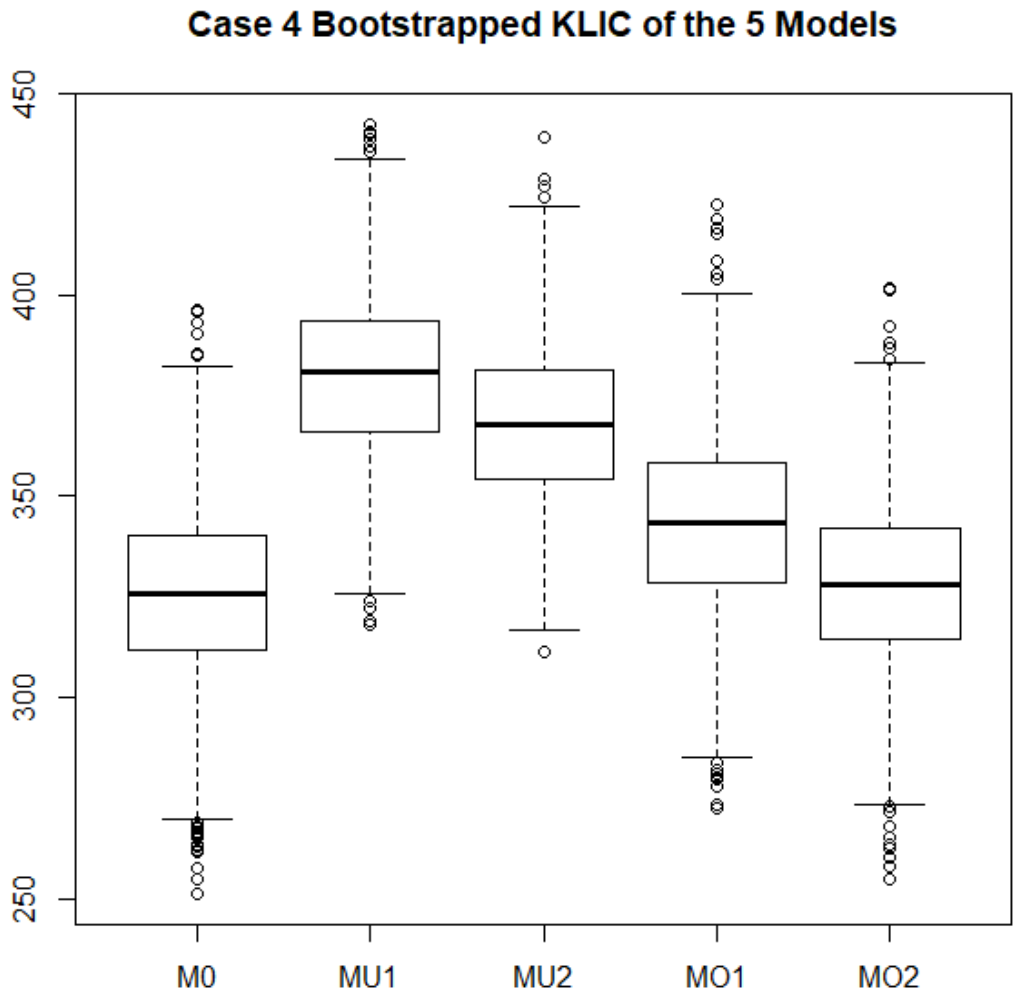

Figure 4. Kullback-Leibler Information Criterion of the 5 Models Estimated in the Simulation for the Large-Sample Continuous Outcome Case (Case 4).

APPENDIX E

R PROGRAMS FOR THE DATA GENERATION AND SIMULATION

R PROGRAMS FOR THE DATA GENERATION AND SIMULATION

```
TDCGen_Ber = function(seed,S,Tvec,rhoyy,rhoxy,TDCTypes,dataTypes,beta,pred)
{set.seed(seed)
seeds = rnorm(S,0,50)
q = length(beta) - 1
beta_x = 3
for(i in 1:S)
{T_i = Tvec[i]
mu <- rep(0, T_i)
y <- rep(0, T_i)
X = matrix(0, T_i, q+1)
px = matrix(0, T_i, q)
X[,1] = rep(1, nrow(X))
set.seed(seeds[i])
seeds_i = rnorm(T_i,0,25)
for(t in 1:T_i)
{set.seed(seeds_i[t])
if(t == 1)
{for(j in 1:q){if(dataTypes[j] == 'c')
{px[t,j] = beta_x
X[t,(j+1)] = rnorm(1, px[t,j], pred[j])}
else{px[t,j] = pred[j]
X[t,(j+1)] = rbinom(1, 1, px[t,j])}}
mu[t] <- exp(X[t,] %*% beta) / (1+exp(X[t,] %*% beta))
y[t] <- rbinom(1, 1, mu[t])}
else{for(j in 1:q){if(dataTypes[j] == 'c')
{if(TDCTypes[j] == 1 | TDCTypes[j] == 2)
{px[t,j] = beta_x
X[t,(j+1)] = rnorm(1, px[t,j], pred[j])}
else{px[t,j] = beta_x
X[t,(j+1)] = rnorm(1, px[t,j], pred[j])}}
else {if(TDCTypes[j] == 1 | TDCTypes[j] == 2)
{px[t,j] = pred[j]
X[t,(j+1)] = rbinom(1, 1, px[t,j])}
else{eta_j = log(pred[j]/(1-pred[j]))
px[t,j] = exp(eta_j) / (1+exp(eta_j))
X[t,(j+1)] = rbinom(1, 1, px[t,j])}}}
types23 = ifelse((TDCTypes>1 & TDCTypes<4), 1, 0)
Xnoint = as.matrix(X[,-1])
values = ifelse((dataTypes=='c'), log((pnorm(Xnoint[t-1,],px[t-1,],pred))/(1-
(pnorm(Xnoint[t-1,],px[t-1,],pred)))),log((px[t-1,]^(Xnoint[t-1,])*(1-px[t-1,])^(1-
Xnoint[t-1,]))/(1-(px[t-1,]^(Xnoint[t-1,]))*(1-px[t-1,])^(1-Xnoint[t-1,]))))
x_prev = t(types23) %*% values
eta_t = X[t,] %*% beta + rhoxy*(x_prev)+rhoyy*log((mu[t-1]^(y[t-1])*(1-mu[t-1])^(1-
y[t-1]))/(1-(mu[t-1]^(y[t-1]))*(1-mu[t-1])^(1-y[t-1])))
```

```
mu[t] <- exp(eta_t) / (1+exp(eta_t))

y[t] <- rbinom(1, 1, mu[t])}}
if(i == 1){Y <- y
Xmat <- X}
else{Y <- c(Y, y)
Xmat <- rbind(Xmat, X)}}
list(yvec = Y, Xmat = Xmat)}

TDCGen_Nor=function(seed,S,Tvec,rhoyy,rhoxy,TDCTypes,dataTypes,beta,pred,sigma)
{set.seed(seed)
seeds = rnorm(S,0,50)
q = length(beta) - 1
beta_x = 3
for(i in 1:S)
{T_i = Tvec[i]
mu <- rep(0, T_i)
y <- rep(0, T_i)
X = matrix(0, T_i, q+1)
px = matrix(0, T_i, q)
X[,1] = rep(1, nrow(X))
set.seed(seeds[i])
seeds_i = rnorm(T_i, 0, 25)
for(t in 1:T_i)
{set.seed(seeds_i[t])
if(t == 1){for(j in 1:q){ if(dataTypes[j] == 'c')
{px[t,j] = beta_x
X[t,(j+1)] = rnorm(1, px[t,j], pred[j])}
else{px[t,j] = pred[j]
X[t,(j+1)] = rbinom(1, 1, px[t,j])}}
mu[t] <- X[t,] %*% beta
y[t] <- rnorm(1, mu[t], sigma)}
else{for(j in 1:q)
{if(dataTypes[j] == 'c'){if(TDCTypes[j] == 1 | TDCTypes[j] == 2)
{px[t,j] = beta_x
X[t,(j+1)] = rnorm(1, px[t,j], pred[j])}
else{px[t,j] = beta_x
X[t,(j+1)] = rnorm(1, px[t,j], pred[j])}}
else{if(TDCTypes[j] == 1 | TDCTypes[j] == 2)
{px[t,j] = pred[j]
X[t,(j+1)] = rbinom(1, 1, px[t,j])}
else{eta_j = log( pred[j]/(1-pred[j]) )
px[t,j] = exp(eta_j)/( 1+exp(eta_j) )
X[t,(j+1)] = rbinom(1, 1, px[t,j])}}}
types23 = ifelse((TDCTypes>1 & TDCTypes<4), 1, 0)
Xnoint = as.matrix(X[,-1])
```

```
values = ifelse((dataTypes=='c'), log((pnorm(Xnoint[t-1,],px[t-1,],pred))/(1-
(pnorm(Xnoint[t-1,],px[t-1,],pred)))),log((px[t-1,]^(Xnoint[t-1,])*(1-px[t-1,])^(1-
Xnoint[t-1,]))/(1-(px[t-1,]^(Xnoint[t-1,]))*(1-px[t-1,])^(1-Xnoint[t-1,]))))
x_prev = t(types23) %*% values
eta_t = X[t,] %*% beta + rhoxy*(x_prev) + rhoyy*log((pnorm(y[t-1],mu[t-1],sigma))/(1-
(pnorm(y[t-1],mu[t-1],sigma))))
mu[t] <- eta_t
y[t] <- rnorm(1, mu[t], sigma)}}
if(i == 1){Y <- y
Xmat <- X}
else{Y <- c(Y, y)
Xmat <- rbind(Xmat, X)}}
list(yvec=Y, Xmat=Xmat)}

GMM_Ber = function(yvec, subjectID, Zmat, Xmat, covTypeVec, Tvec, N)
{if(!is.matrix(Zmat)){K0 = 0}
else if(is.matrix(Zmat)){K0 = ncol(Zmat)}
Ktv = ncol(Xmat)
K = 1+K0+Ktv
Tmax = max(Tvec)
K1 = 0
K2 = 0
K3 = 0
for(k in 1:Ktv)
{if(covTypeVec[k]==1)
{K1 = K1+1}
else{ if (covTypeVec[k]==2)
{K2 = K2+1}
else {K3 = K3+1}}}
Lmax = 1*Tmax + K0*Tmax  + (Tmax^2)*K1 + Tmax*(Tmax+1)/2*K2 + Tmax*K3
if(K0==0){ZX = Xmat}
else if(K0!=0){ZX = cbind(Zmat,Xmat)}
betaI = gee(yvec~ZX,id=subjectID,family=binomial,corstr="independence")$coefficients
#betaI = c(rep(0,K))
QuadForm = function(beta)
{G = rep(0,Lmax)
VN = matrix(0,Lmax,Lmax)
Count = rep(0,Lmax)
for(i in 1:N)
{subjectIndex = sum(Tvec[0:(i-1)])+1
Est_i = ValidMomentCalc_Ber(yvec, subjectIndex, Zmat, Xmat, covTypeVec, beta,
Tvec[i], Tmax, Count)
gEst_i = Est_i[[1]]
Count = Est_i[[2]]
G = G + gEst_i
VN = VN + gEst_i%*%t(gEst_i)}
```

```
G = G / Count
D = matrix(0,Lmax,Lmax)
for(i in 1:Lmax)
{for(j in 1:Lmax)
{D[i,j] = min(Count[i],Count[j])}}
W = ginv(VN / D)
QF = t(G) %*% W %*% G
QF}
betahat = optim(betaI, QuadForm)$par
dBetaG = matrix(0,Lmax,K)
VN = matrix(0,Lmax,Lmax)
Count = rep(0,Lmax)
for(i in 1:N)
{subjectIndex = sum(Tvec[0:(i-1)])+1
Est_i = ValidMomentCalc_Ber(yvec, subjectIndex, Zmat, Xmat, covTypeVec, betahat,
Tvec[i], Tmax, Count)
gEst_i = Est_i[[1]]
Count = Est_i[[2]]
VN = VN + gEst_i%*%t(gEst_i)
dBetagEst_i = ValidMomentDeriv_Ber(yvec, subjectIndex, Zmat, Xmat, covTypeVec,
betahat, Tvec[i], Tmax)
dBetaG = dBetaG + dBetagEst_i}
D = matrix(0,Lmax,Lmax);
for(i in 1:Lmax)
{for(j in 1:Lmax)
{D[i,j] = min(Count[i],Count[j])}}
Divisor = matrix(c(rep(Count,K)),length(Count),K)
dBetaG = dBetaG / Divisor
W = ginv(VN / D)
AsymptoticWeight = t(dBetaG) %*% W %*% dBetaG
AsymptoticCovariance = (1/N)*ginv(AsymptoticWeight)
list(betaHat=betahat, covEst = AsymptoticCovariance)}

ValidMomentCalc_Ber = function(yvec, subjectIndex, Zmat, Xmat, covTypeVec, betaI,
T, Tmax, Count)
{if(!is.matrix(Zmat)){K0 = 0}
else if(is.matrix(Zmat)){K0 = ncol(Zmat)}
Ktv = ncol(Xmat)
K = 1+K0+Ktv
K1 = 0
K2 = 0
K3 = 0
for(k in 1:Ktv)
{if(covTypeVec[k]==1)
{K1 = K1+1}
else{if (covTypeVec[k]==2)
```

```
{K2 = K2+1}
else {K3 = K3+1} }}
Lmax = 1*Tmax + K0*Tmax  + (Tmax^2)*K1 + Tmax*(Tmax+1)/2*K2 + Tmax*K3
yvec_i = yvec[subjectIndex:(subjectIndex+T)]
mu_i = rep(0,T)
eta_i = rep(0,T)
for(t in 1:T)
{if(K0!=0){zmat_it = Zmat[subjectIndex+t-1,]}
xmat_it = Xmat[subjectIndex+t-1,]
if(K0==0){zx_it = c(1,xmat_it)}
else if(K0!=0){zx_it = c(1,zmat_it,xmat_it)}
eta_i[t] = zx_it %*% betaI
mu_i[t] = exp(eta_i[t])/(1+exp(eta_i[t]))}
gEst_i = rep(0,Lmax)
count = 1
for(t in 1:T)
{gEst_i[count] = (mu_i[t]/(1+exp(eta_i[t])))*(yvec_i[t]-mu_i[t])
Count[count] = Count[count]+1
count = count+1}
count = count + (Tmax-T)
if(K0!=0)
{for(k in 1:K0)
{for(t in 1:T)
{gEst_i[count] = (mu_i[t]/(1+exp(eta_i[t])))*Zmat[subjectIndex+t-1,k]*(yvec_i[t]-
mu_i[t])
Count[count] = Count[count]+1
count = count+1}
count = count + (Tmax-T)}}
for (k in 1:Ktv)
{if(covTypeVec[k]==1)
{for (s in 1:T)
{for (t in 1:T)
{gEst_i[count] = (mu_i[s]/(1+exp(eta_i[s])))*Xmat[subjectIndex+s-1,k]*(yvec_i[t]-
mu_i[t])
Count[count] = Count[count]+1
count = count + 1}
count = count + (Tmax-T)}
count = count + Tmax*(Tmax-T)}
else if(covTypeVec[k]==2)
{for (s in 1:T)
{for (t in 1:s)
{gEst_i[count] = (mu_i[s]/(1+exp(eta_i[s])))*Xmat[subjectIndex+s-1,k]*(yvec_i[t]-
mu_i[t])
Count[count] = Count[count]+1
count = count + 1}}
count = count + (1/2)*(Tmax*(Tmax+1)-T*(T+1))}
```

```
else{for (s in 1:T)
{gEst_i[count] = (mu_i[s]/(1+exp(eta_i[s])))*Xmat[subjectIndex+s-1,k]*(yvec_i[s]-
mu_i[s])
Count[count] = Count[count]+1
count = count + 1}
count = count + (Tmax-T)}}
list(gEst_i,Count)}

ValidMomentDeriv_Ber = function(yvec, subjectIndex, Zmat, Xmat, covTypeVec, betaI,
T, Tmax)
{if(!is.matrix(Zmat)){K0 = 0}
else if(is.matrix(Zmat)){K0 = ncol(Zmat)}
Ktv = ncol(Xmat)
K = 1+K0+Ktv
K1 = 0
K2 = 0
K3 = 0
for(k in 1:Ktv)
{if(covTypeVec[k]==1){K1 = K1+1}
else{if (covTypeVec[k]==2){K2 = K2+1}
else {K3 = K3+1} }}
Lmax = 1*Tmax + K0*Tmax  + (Tmax^2)*K1 + Tmax*(Tmax+1)/2*K2 + Tmax*K3
yvec_i = yvec[subjectIndex:(subjectIndex+T)]
mu_i = rep(0,T)
eta_i = rep(0,T)
for(t in 1:T)
{if(K0!=0){zmat_it = Zmat[subjectIndex+t-1,]}
xmat_it = Xmat[subjectIndex+t-1,]
if(K0==0){zx_it = c(1,xmat_it)}
else if(K0!=0){zx_it = c(1,zmat_it,xmat_it)}
eta_i[t] = zx_it %*% betaI
mu_i[t] = exp(eta_i[t])/(1+exp(eta_i[t]))}
dBetamu_i = matrix(0,T,K)
for(t in 1:T)
{dCount = 1
dBetamu_i[t,dCount] = (1)*mu_i[t]*(1-mu_i[t])
dCount = dCount+1
if(K0!=0)
{for(j in 1:K0)
{dBetamu_i[t,dCount] = (Zmat[subjectIndex+t-1,j])*mu_i[t]*(1-mu_i[t])
dCount = dCount+1}}
for(j in 1:ncol(Xmat))
{dBetamu_i[t,dCount] = (Xmat[subjectIndex+t-1,j])*mu_i[t]*(1-mu_i[t])
dCount = dCount+1}}
d2Betamu_i_part = matrix(0,T,K)
for (t in 1:T)
```

```
{for (k in 1:K)
{d2Betamu_i_part[t,k] = dBetamu_i[t,k]*(1-2*mu_i[t])}}
dBetag_i = matrix(0,Lmax,K)
count = 1
for(t in 1:T)
{s = t
j = 1
for(k in 1:K)
{dBetag_i[count,k] = (-1)*dBetamu_i[s,j]*dBetamu_i[t,k] + (1)*
d2Betamu_i_part[s,k]*(yvec_i[t]-mu_i[t])}
count = count+1}
count = count + (Tmax-T)
if(K0!=0)
{for(j in 1:K0)
{for(t in 1:T)
{s=t
for(k in 1:K)
{dBetag_i[count,k] = (-1)*dBetamu_i[s,1+j]*dBetamu_i[t,k] + (Zmat[(subjectIndex+s-
1),j])*d2Betamu_i_part[s,k]*(yvec_i[t]-mu_i[t])}
count = count+1}
count = count + (Tmax-T)}}
for (j in 1:Ktv)
{if(covTypeVec[j]==1)
{for (s in 1:T)
{for (t in 1:T)
{for(k in 1:K)
{dBetag_i[count,k] = (-1)*dBetamu_i[s,1+K0+j]*dBetamu_i[t,k] +
(Xmat[(subjectIndex+s-1),j])*d2Betamu_i_part[s,k]*(yvec_i[t]-mu_i[t])}
count = count + 1}
count = count + (Tmax-T)}
count = count + Tmax*(Tmax-T)}
else if(covTypeVec[j]==2)
{for (s in 1:T)
{for (t in 1:s)
{for(k in 1:K)
{dBetag_i[count,k] = (-1)*dBetamu_i[s,1+K0+j]*dBetamu_i[t,k] +
(Xmat[(subjectIndex+s-1),j])*d2Betamu_i_part[s,k]*(yvec_i[t]-mu_i[t])}
count = count + 1}}
count = count + (1/2)*(Tmax*(Tmax+1)-T*(T+1))}
else{for (s in 1:T)
{for(k in 1:K)
{dBetag_i[count,k] = (-1)*dBetamu_i[s,1+K0+j]*dBetamu_i[t,k] +
(Xmat[(subjectIndex+s-1),j])*d2Betamu_i_part[s,k]*(yvec_i[t]-mu_i[t])}
count = count + 1}
count = count + (Tmax-T)}}
dBetag_i}
```

```
T=5
# Small Bin
BinSmallKLICout=matrix(ncol=6,nrow=2000)
for (i in 1 : 2000)
{S = 100
Tvec = rep(T, S)
BinarySmall <- TDCGen_Ber(seed = 12345+i, S = 100, Tvec = rep(T, S), rhoyy = 0.25,
rhoxy = 0.25, TDCTypes <- c(1,2,3,2,3), dataTypes <- c("c","c","c","c","c"), beta <-
c(0.58, -0.049, -0.01, -0.091, -0.280, 0.004), pred <- c(0.1852, 0.0216, 0.000028, 0.0025,
0.0092, 0.0095) )
BinSmY <- matrix(BinarySmall$yvec, nrow=S, ncol=T, byrow=FALSE)
BinarySmall$Xmat[,1] = BinarySmall$yvec
BS = as.data.frame(BinarySmall$Xmat)
names(BS) <- c('y','x1','x2','x3','x4','x5')
full=GMM_Ber(BinarySmall$yvec, seq(1:100, 5), rep(0,ncol(BS)-1), BS[,-c(1,4,6)],
c(1,2,3), rep(T, S), 100)
lackX2=GMM_Ber(BinarySmall$yvec, seq(1:100, 5), rep(0,ncol(BS)-1), BS[,-
c(1,3,4,6)], c(1,3), rep(T, S), 100)
lackX4=GMM_Ber(BinarySmall$yvec, seq(1:100, 5), rep(0,ncol(BS)-1), BS[,-
c(1,4,5,6)], c(1,2), rep(T, S), 100)
withX3=GMM_Ber(BinarySmall$yvec, seq(1:100, 5), rep(0,ncol(BS)-1), BS[,-c(1,6)],
c(1,2,3,2), rep(T, S), 100)
withX5=GMM_Ber(BinarySmall$yvec, seq(1:100, 5), rep(0,ncol(BS)-1), BS[,-c(1,4)],
c(1,2,2,3), rep(T, S), 100)
All=GMM_Ber(BinarySmall$yvec, seq(1:100, 5), rep(0,ncol(BS)-1), BS[,-1],
c(1,2,3,2,3), rep(T, S), 100)
a = klic(full)
b = klic(lackX2)
c = klic(lackX4)
d = klic(withX3)
e = klic(withX5)
f = klic(All)
out <- c(a,b,c,d,e,f)
BinSmallKLICout[i,] = out}
BinSmallKLIC = as.data.frame(BinSmallKLICout)
names(BinSmallKLICout) <- c('full','lackX2','lackX4','withX3','withX5','All')
head(BinSmallKLICout)
write.csv(BinSmallKLIC, file="BinSmall.csv")

# Large Bin
BinLargeKLIC= matrix(ncol=6,nrow=2000)
for (i in 1 : 2000)
{S = 500
Tvec = rep(T, S)
BinaryLarge <- TDCGen_Ber(seed = 12345, S = 500, Tvec = rep(T, S), rhoyy = 0.25,
```

```
rhoxy = 0.25, TDCTypes <- c(1,2,3,2,3), dataTypes <- c("c","c","c","c","c"), beta <-
c(0.58, -0.049, -0.01, -0.091, -0.280, 0.004), pred <- c(0.1852, 0.0216, 0.000028, 0.0025,
0.0092, 0.0095) )
BinLargeY <- matrix(BinaryLarge$yvec, nrow=S, ncol=T, byrow=FALSE)
BinaryLarge$Xmat[,1] = BinaryLarge$yvec
BL = as.data.frame(BinaryLarge$Xmat)
names(BL) <- c('y','x1','x2','x3','x4','x5')
full=GMM_Ber(BinaryLarge$yvec, seq(1:500, 5), rep(0,ncol(BL)-1), BL[,-c(1,4,6)],
c(1,2,3), rep(T, S), 500)
lackX2=GMM_Ber(BinaryLarge$yvec, seq(1:500, 5), rep(0,ncol(BL) -1), BL[,-
c(1,3,4,6)], c(1,3), rep(T, S), 500)
lackX4=GMM_Ber(BinaryLarge$yvec, seq(1:500, 5), rep(0,ncol(BL) -1), BL[,-
c(1,4,5,6)], c(1,2), rep(T, S), 500)
withX3=GMM_Ber(BinaryLarge$yvec, seq(1:500, 5), rep(0,ncol(BL) -1), BL[,-c(1,6)],
c(1,2,3,2), rep(T, S), 500)
withX5=GMM_Ber(BinaryLarge$yvec, seq(1:500, 5), rep(0,ncol(BL) -1), BL[,-c(1,4)],
c(1,2,2,3), rep(T, S), 500)
All=GMM_Ber(BinaryLarge$yvec, seq(1:500, 5), rep(0,ncol(BL) -1), BL[,-1],
c(1,2,3,2,3), rep(T, S), 500)
a = klic(full)
b = klic(lackX2)
c = klic(lackX4)
d = klic(withX3)
e = klic(withX5)
f = klic(All)
out <- c(a,b,c,d,e,f)
BinLargeKLIC [i,] = out}
BinLargeKLIC = as.data.frame(BinLargeKLIC)
names(BinLargeKLIC) <- c('full','lackX2','lackX4','withX3','withX5','All')
write.csv(BinLargeKLIC, file="BinLarge.csv")

# Small Cont
ContSmallKLIC = matrix(ncol=6,nrow=2000)
for (i in 1 : 2000)
{S = 100
Tvec = rep(T, S)
ContSmall <- TDCGen_Nor(seed = 12345, S = 100, Tvec = rep(T, S), rhoyy = 0.25,
rhoxy = 0.25, TDCTypes <- c(1,2,3,2,3), dataTypes <- c("c","c","c","c","c"), beta <-
c(0.58, -0.049, -0.01, -0.091, -0.280, 0.004), pred <- c(0.1852, 0.0216, 0.000028, 0.0025,
0.0092, 0.0095), sigma <- c(1, 2.2, 3.5, 1.5, 4.2, 0.8) )
ContSmY <- matrix(ContSmall$yvec, nrow=S, ncol=T, byrow=FALSE)
ContSmall$Xmat[,1] = ContSmall$yvec
CS = as.data.frame(ContSmall$Xmat)
names(CS) <- c('y','x1','x2','x3','x4','x5')
full=GMM_Nor(ContSmall$yvec, seq(1:100, 5), rep(0,ncol(CS)-1), CS [,-c(1,4,6)],
c(1,2,3), rep(T, S), 100)
```

```
lackX2=GMM_Nor(ContSmall$yvec, seq(1:100, 5), rep(0,ncol(CS)-1), CS [,-c(1,3,4,6)],
c(1,3), rep(T, S), 100)
lackX4=GMM_Nor(ContSmall$yvec, seq(1:100, 5), rep(0,ncol(CS)-1), CS [,-c(1,4,5,6)],
c(1,2), rep(T, S), 100)
withX3=GMM_Nor(ContSmall$yvec, seq(1:100, 5), rep(0,ncol(CS)-1), CS [,-c(1,6)],
c(1,2,3,2), rep(T, S), 100)
withX5=GMM_Nor(ContSmall$yvec, seq(1:100, 5), rep(0,ncol(CS)-1), CS [,-c(1,4)],
c(1,2,2,3), rep(T, S), 100)
All=GMM_Nor(ContSmall$yvec, seq(1:100, 5), rep(0,ncol(CS)-1), CS [,-1], c(1,2,3,2,3),
rep(T, S), 100)
a = klic(full)
b = klic(lackX2)
c = klic(lackX4)
d = klic(withX3)
e = klic(withX5)
f = klic(All)
out <- c(a,b,c,d,e,f)
ContSmallKLIC[i,] = out}
ContSmallKLIC = as.data.frame(ContSmallKLIC)
names(ContSmallKLIC) <- c('full','lackX2','lackX4','withX3','withX5','All')
write.csv(ContSmallKLIC, file="ContSmall.csv")

# Large Cont
ContLargeKLIC = matrix(ncol=6,nrow=2000)
for (i in 1 : 2000)
{S = 500
Tvec = rep(T, S)
ContLarge <- TDCGen_Nor(seed = 12345, S = 500, Tvec = rep(T, S), rhoyy = 0.25,
rhoxy = 0.25, TDCTypes <- c(1,2,3,2,3), dataTypes <- c("c","c","c","c","c"), beta <-
c(0.58, -0.049, -0.01, -0.091, -0.280, 0.004), pred <- c(0.1852, 0.0216, 0.000028, 0.0025,
0.0092, 0.0095), sigma <- c(1, 2.2, 3.5, 1.5, 4.2, 0.8) )
ContLargeY <- matrix(ContLarge$yvec, nrow=S, ncol=T, byrow=FALSE)
ContLarge$Xmat[,1] = ContLarge$yvec
CL = as.data.frame(ContLarge$Xmat)
names(CL) <- c('y','x1','x2','x3','x4','x5')
full=GMM_Nor(ContLarge$yvec, seq(1:500, 5), rep(0,ncol(CL)-1), CL[,-c(1,4,6)],
c(1,2,3), rep(T, S), 500)
lackX2=GMM_Nor(ContLarge$yvec, seq(1:500, 5), rep(0,ncol(CL) -1), CL[,-c(1,3,4,6)],
c(1,3), rep(T, S), 500)
lackX4=GMM_Nor(ContLarge$yvec, seq(1:500, 5), rep(0,ncol(CL) -1), CL[,-c(1,4,5,6)],
c(1,2), rep(T, S), 500)
withX3=GMM_Nor(ContLarge$yvec, seq(1:500, 5), rep(0,ncol(CL) -1), CL[,-c(1,6)],
c(1,2,3,2), rep(T, S), 500)
withX5=GMM_Nor(ContLarge$yvec, seq(1:500, 5), rep(0,ncol(CL) -1), CL[,-c(1,4)],
c(1,2,2,3), rep(T, S), 500)
```

```
All=GMM_Nor(ContLarge$yvec, seq(1:500, 5), rep(0,ncol(CL) -1), CL[,-1], c(1,2,3,2,3),
rep(T, S), 500)
a = klic(full)
b = klic(lackX2)
c = klic(lackX4)
d = klic(withX3)
e = klic(withX5)
f = klic(All)
out <- c(a,b,c,d,e,f)
ContLargeKLIC[i,] = out}
ContLargeKLIC = as.data.frame(ContLargeKLIC)
names(ContLargeKLIC) <- c('full','lackX2','lackX4','withX3','withX5','All')
write.csv(ContLargeKLIC, file="ContLarge.csv")
out1 <- colMeans(BinSmallKLIC)
out2 <- colMeans(BinLargeKLIC)
out3 <- colMeans(ContSmallKLIC)
out4 <- colMeans(ContLargeKLIC)
outs <- cbind(out1,out2,out3,out4)

BootBinSmall=matrix(ncol=5,nrow=1000)
colnames(BootBinSmall)=c('M0','Mu1','Mu2','Mo1','Mo2')
for (i in 1 : 1000){for (j in 1 : 5){
BootBinSmall[i,j]=mean(sample(BinSmallKLIC[,j],size=1000,replace=TRUE))}}
boxplot(BootBinSmall,las=2,horizontal=TRUE,notch=TRUE)

BootBinLarge=matrix(ncol=5,nrow=1000)
colnames(BootBinLarge)=c('M0','Mu1','Mu2','Mo1','Mo2')
for (i in 1 : 1000){for (j in 1 : 5){
BootBinLarge[i,j]=mean(sample(BinLargeKLIC[,j],size=1000,replace=TRUE))}}
boxplot(BootBinLarge,las=2,horizontal=TRUE,notch=TRUE)

BootContSmall=matrix(ncol=5,nrow=1000)
colnames(BootContSmall)=c('M0','Mu1','Mu2','Mo1','Mo2')
for (i in 1 : 1000){for (j in 1 : 5){
BootContSmall[i,j]=mean(sample(ContSmallKLIC[,j],size=1000,replace=TRUE))}}
boxplot(BootContSmall,las=2,horizontal=TRUE,notch=TRUE)

BootContLarge=matrix(ncol=5,nrow=1000)
colnames(BootContLarge)=c('M0','Mu1','Mu2','Mo1','Mo2')
for (i in 1 : 1000){for (j in 1 : 5){
BootContLarge[i,j]=mean(sample(ContLargeKLIC[,j],size=1000,replace=TRUE))}}
boxplot(BootContLarge,las=2,horizontal=TRUE,notch=TRUE)
```